Vitória Rodrigues Pinto Borelli Figueiredo

Geração e Avaliação de Dados Sintéticos para Aplicações em Saúde Mental

São Carlos, SP

2026

Vitória Rodrigues Pinto Borelli Figueiredo

Geração e Avaliação de Dados Sintéticos para Aplicações em Saúde Mental

Trabalho de Conclusão de Curso apresentado ao curso de Engenharia de Computação da Universidade Federal de São Carlos, como requisito parcial para a obtenção do título de Bacharel em Engenharia de Computação.

Universidade Federal de São Carlos – UFSCar

Departamento de Computação

Programa de Graduação

Orientador: Helena de Medeiros Caseli

São Carlos, SP

2026

Dedico este trabalho aos meus pais, que sempre acreditaram em mim, mesmo quando eu não o fiz.

Agradecimentos

Os agradecimentos principais são direcionados à minha professora e orientadora Dra. Helena Caseli, que, sempre com muita paciência e leveza, me guiou ao longo do desenvolvimento deste trabalho. Sua solicitude e didática são exemplares e foram essenciais para a obtenção dos resultados aqui apresentados. Seu entusiasmo e paixão pela área são verdadeiramente inspiradores.

Agradeço também a todos que estiveram ao meu lado, não apenas neste estágio final, mas ao longo de toda a graduação. Aos professores que, cada um à sua maneira, nos incentivaram não somente a aprender o conteúdo, mas a sermos questionadores. Sou grata a todos os educadores, incluindo aqueles que me formaram antes da universidade, que nos instigam e nos propõem desafios que, muitas vezes desconfortáveis, são justamente os que mais nos transformam.

Além deles, é impossível não mencionar os queridos amigos que fiz ao longo da graduação. Esse período certamente não teria sido o mesmo sem eles, que se tornaram minha base em São Carlos. Muito além de parceiros de estudos e trabalhos, foram companheiros de risadas e de dias difíceis, tornando o caminho mais leve e prazeroso. Com eles, nenhum dia foi monótono, e são lembranças que carrego com muito afeto.

E, claro, não poderia deixar de mencionar meus maiores incentivadores: meus pais, Marcelo e Adriane. Sempre presentes, sempre dispostos a me ouvir e a me acolher, foram eles que me deram coragem para seguir em frente nos momentos em que o caminho parecia mais incerto. Me deram liberdade para escolher meu próprio caminho e, ao mesmo tempo, a segurança de saber que nunca estaria sozinha. Todo o carinho e apoio que sempre me ofereceram me permitiram chegar até aqui.

Por fim, gostaria também de dedicar um agradecimento especial à minha avó Marlene, que sempre acendeu em mim a curiosidade e o gosto por explorar e aprender. Ela costumava dizer que o conhecimento é a única coisa que ninguém nos pode tirar, e esse ensinamento se tornou um valor que carrego comigo até hoje.

Este trabalho foi desenvolvido no escopo do projeto “Suporte baseado em IA para comunicação em saúde mental” (AIM-Health) apoiado pela FAPESP #24/10233-7.

"A melhor forma de predizer o futuro é inventá-lo."
(Alan Kay)

Resumo

Corpora anotados na área de saúde mental são escassos e frequentemente não podem ser redistribuídos publicamente devido a restrições éticas e legais associadas à sensibilidade dos dados, o que limita a reprodutibilidade de experimentos e o avanço colaborativo da área. Este trabalho propõe e avalia um pipeline de geração de dados sintéticos para o *corpus* Amive, composto por postagens anônimas de redes sociais anotadas com sintomas de depressão, com o objetivo de viabilizar a distribuição pública de um *corpus* funcional sem expor os textos originais dos usuários que os produziram. O pipeline combina tradução automática via modelo de linguagem (português para inglês e retradução ao português), geração de paráfrases com o modelo PEGASUS por meio de *diverse beam search*, e seleção da melhor paráfrase candidata com base na similaridade semântica calculada pelo BERTScore, utilizando o BERTimbau como modelo de representação. A avaliação do *corpus* sintético gerado, com foco no sintoma de Tristeza/Humor depressivo, é conduzida em duas frentes complementares: uma avaliação extrínseca, na qual classificadores binários de diferentes naturezas (Regressão Logística, SVM, Naive Bayes e Random Forest) são treinados e testados em quatro cenários combinando dados originais e sintéticos; e uma avaliação intrínseca, baseada em métricas de diversidade lexical e na distribuição de categorias gramaticais do *corpus* gerado, complementada por uma análise qualitativa de exemplos representativos. Os resultados indicam que o *corpus* sintético preserva, em grande medida, a utilidade do *corpus* original para a tarefa de classificação, com destaque para a revocação, métrica considerada mais crítica em aplicações de triagem em saúde mental, na qual deixar de identificar um caso positivo tende a ser mais custoso do que um alarme falso. Observa-se, no entanto, queda mais consistente na precisão dos classificadores treinados exclusivamente com dados sintéticos quando avaliados sobre dados reais, além de indícios de que parte da diversidade lexical introduzida decorre de uma tendência do pipeline a podar conteúdo em sentenças sintaticamente complexas, e não apenas de reformulação genuína. A análise qualitativa revela, ainda, padrões recorrentes de perda de nuances discursivas, como despersonalização de relatos autobiográficos e literalização de construções metafóricas. Conclui-se que o pipeline proposto constitui um mecanismo viável, embora não isento de limitações, para a geração de *corpora* sintéticos em português no domínio da saúde mental, oferecendo uma alternativa concreta à indisponibilidade de dados anotados nesse domínio.

Palavras-chave: Dados sintéticos. Processamento de Linguagem Natural. Saúde mental. Privacidade de dados. Paráfrase automática.

Abstract

Annotated corpora in the mental health domain are scarce and frequently cannot be publicly redistributed due to ethical and legal restrictions associated with the sensitivity of the data involved, which limits the reproducibility of experiments and collaborative progress in the field. This work proposes and evaluates a synthetic data generation pipeline for the Amive corpus, composed of anonymous social media posts annotated with depression symptoms, aiming to enable the public distribution of a functional corpus without exposing the original texts of the users who produced them. The pipeline combines machine translation via a language model (Portuguese to English and back-translation to Portuguese), paraphrase generation with the PEGASUS model through diverse beam search, and selection of the best candidate paraphrase based on semantic similarity computed by BERTScore, using BERTimbau as the representation model. The evaluation of the generated synthetic corpus, focused on the Sadness/Depressed Mood symptom, is conducted along two complementary lines: an extrinsic evaluation, in which binary classifiers of different natures (Logistic Regression, SVM, Naive Bayes, and Random Forest) are trained and tested across four scenarios combining original and synthetic data; and an intrinsic evaluation, based on lexical diversity metrics and the part-of-speech distribution of the generated corpus, complemented by a qualitative analysis of representative examples. Results indicate that the synthetic corpus largely preserves the utility of the original corpus for the classification task, with particular emphasis on recall, a metric considered more critical in mental health screening applications, where failing to identify a positive case tends to be more costly than a false alarm. However, a more consistent drop in precision is observed for classifiers trained exclusively on synthetic data when evaluated on real data, along with evidence that part of the lexical diversity introduced stems from a tendency of the pipeline to prune content in syntactically complex sentences, rather than from genuine reformulation alone. The qualitative analysis further reveals recurring patterns of discursive nuance loss, such as the depersonalization of autobiographical accounts and the literalization of metaphorical constructions. It is concluded that the proposed pipeline constitutes a viable, though not limitation-free, mechanism for generating synthetic corpora in Portuguese within the mental health domain, offering a concrete alternative to the unavailability of annotated data in this field.

Keywords: Synthetic data. Natural Language Processing. Mental health. Data privacy. Automatic paraphrasing.

Lista de ilustrações

Figura 1 – Arquitetura Transformer. O pontilhado violeta representa o codificador e o pontilhado verde representa o decodificador. 25
Figura 2 – Representações distribuídas: vetores de palavras e sua visualização em espaço bidimensional após redução de dimensionalidade. 26
Figura 3 – Representação do mecanismo de atenção simples (esquerda) e mecanismo de atenção multi-cabeça (direita) para a sentença “The animal didn’t cross the street because it was too tired”. 28
Figura 4 – Matriz de confusão 30
Figura 5 – *Pipeline* experimental: geração do *dataset* sintético e avaliação por classificação binária. 45
Figura 6 – Comparação dos resultados por métrica entre os diferentes cenários avaliados 62
Figura 7 – Comparação dos resultados por modelo entre os diferentes cenários avaliados 64
Figura 8 – Mapas de calor das diferenças de desempenho em relação ao conjunto original 65
Figura 9 – Distribuição de classes gramaticais do *corpus* sintético comparado ao original **(classe positiva)** 68
Figura 10 – Distribuição de classes gramaticais do *corpus* sintético comparado ao original **(classe de controle)** 68
Figura 11 – Comparação das nuvens de palavras do *corpus* original e do *corpus* sintético (**classe positiva**) 73
Figura 12 – Comparação das nuvens de palavras do *corpus* original e do *corpus* sintético (**classe de controle**) 74
Figura 13 – Fluxograma do processo de tradução automática via LLM. 83
Figura 14 – Fluxograma do processo de geração de paráfrases com PEGASUS e *diverse beam search*. 84
Figura 15 – Fluxo de decisão para filtragem e seleção da paráfrase via BERTScore. 85

Lista de quadros

Quadro 1 – Exemplo de preservação semântica adequada (*score* = 73,0) 69
Quadro 2 – Exemplo de preservação parcial (*score* = 71,02) 70
Quadro 3 – Exemplo de perda semântica crítica (*score* = 41,5) 70
Quadro 4 – Exemplo de despersonalização do relato (*score* = 72,71) 71

Lista de tabelas

Tabela 1 – Comparativo dos trabalhos relacionados 42
Tabela 2 – Distribuição de instâncias anotadas por sintoma no *corpus* Amive (WILKENS et al., 2026) 51
Tabela 3 – Estatísticas do processo de geração do corpus sintético 56
Tabela 4 – Resultados dos classificadores para o sintoma Tristeza/Humor depressivo – Acurácia e Precisão 61
Tabela 5 – Resultados dos classificadores para o sintoma Tristeza/Humor depressivo – Recall e F1 61
Tabela 6 – Métricas de qualidade do *corpus* sintético comparado ao original **(classe positiva)**. 66
Tabela 7 – Métricas de qualidade do *corpus* sintético comparado ao original **(classe de controle)**. 66

Lista de abreviaturas e siglas

PLN	Processamento de Linguagem Natural
RNN	Rede Neural Recorrente (*Recurrent Neural Network*)
LSTM	Rede de Memória de Longo Prazo (*Long Short-Term Memory*)
NMT	Tradução Neural de Máquina (*Neural Machine Translation*)
BLEU	*Bilingual Evaluation Understudy*
LDP	Privacidade Diferencial Local (*Local Differential Privacy*)
GAN	Rede Generativa Adversarial (*Generative Adversarial Network*)
LLM	*Large Language Model*
SLM	*Small Language Model*
GSG	*Gap Sentence Generation*
IAA	*Inter-Annotator Agreement*
TF-IDF	*Term Frequency–Inverse Document Frequency*
SVM	*Support Vector Machine*
PPD	Possível Perfil Depressivo

Sumário

1 INTRODUÇÃO 21

2 FUNDAMENTAÇÃO TEÓRICA 23
2.1 Conceitos de Privacidade em PLN 23
2.1.1 Dados Sensíveis 23
2.1.2 Desidentificação e Reidentificação 23
2.2 Arquitetura Transformer 24
2.2.1 Estrutura Geral 25
2.2.2 Representações Distribuídas 26
2.2.3 Codificação Posicional 27
2.2.4 Mecanismo de Atenção 27
2.2.5 Atenção Multi-Cabeça 28
2.2.6 Atenção no Decoder 28
2.3 Tradução Automática Neural 29
2.4 Métricas de Avaliação 30
2.4.1 Métricas de Classificação 30
2.4.2 Métricas de Propriedades Textuais 32

3 TRABALHOS RELACIONADOS 37
3.1 PLN Aplicado à Saúde Mental 37
3.2 Privacidade e a Necessidade de Dados Sintéticos 38
3.3 Geração de Dados Sintéticos em Texto 39
3.4 Avaliação de Dados Sintéticos 41
3.5 Síntese 42

4 *PIPELINE* PARA GERAÇÃO DE DADOS SINTÉTICOS 45
4.1 Métodos 45
4.1.1 Geração do *Dataset* Sintético (Etapa 1) 45
4.1.2 Classificação e Avaliação (Etapas 2 e 3) 47
4.2 Materiais 50
4.2.1 *Corpus* Amive 50
4.2.2 Ferramentas 52

5 EXPERIMENTOS E RESULTADOS 53
5.1 Geração dos Dados Sintéticos 53
5.1.1 *Back-Translation* 53

5.1.2 Geração de Paráfrases . . . 53
5.1.3 Processamento dos Resultados . . . 55
5.1.4 Estatísticas do Processo de Geração . . . 56
5.2 Treinamento dos Classificadores . . . 58
5.2.1 Construção e Organização dos *Datasets* . . . 58
5.2.2 Vetorização e Classificação . . . 59
5.3 Avaliação dos Classificadores . . . 60
5.3.1 Análise por Métrica . . . 60
5.3.2 Análise por Modelo . . . 62
5.3.3 Mapa de Calor . . . 63
5.3.4 Síntese . . . 64
5.4 Avaliação Intrínseca . . . 65
5.4.1 Vocabulário, D-2 e BLEU . . . 65
5.4.2 Distribuição de categorias gramaticais . . . 66
5.5 Discussão dos Resultados . . . 67
5.5.1 Análise Qualitativa . . . 67
5.5.2 Discussão Conjunta dos Resultados . . . 75

6 CONCLUSÕES . . . 79
6.1 Limitações . . . 79
6.2 Trabalhos Futuros . . . 80

APÊNDICE A – DIAGRAMAS DETALHADOS DO *PIPELINE* DE GERAÇÃO DE DADOS SINTÉTICOS . . . 83

REFERÊNCIAS . . . 87

1 Introdução

Transtornos de saúde mental como depressão e ansiedade afetam uma parcela crescente da população mundial, e a identificação precoce de seus sinais representa um desafio clínico relevante. Com a popularização das redes sociais, o conteúdo publicado por usuários passou a ser explorado como fonte de dados para tarefas de triagem automática em Processamento de Linguagem Natural (PLN) (SANTOS; OLIVEIRA; PARABONI, 2024; CHOUDHURY et al., 2013; ALHAMED; IVE; SPECIA, 2024).

No entanto, *corpora* anotados nesse domínio, isto é, coleções de textos utilizados para análise e treinamento de modelos, rotulados com informações clínicas, são difíceis de se obter e ainda escassos no que diz respeito à sua distribuição. Como todo processo manual, a anotação de dados neste domínio requer tempo e recursos, o que dificulta a construção de grandes *datasets* para treinamento de modelos de PLN (PAGANO et al., 2026). Principalmente, postagens em redes sociais carregam informações sensíveis sobre saúde mental de indivíduos identificáveis, o que impõe barreiras éticas e legais ao seu compartilhamento (PARABONI; CASELI, 2024). Na prática, *corpora* construídos nesse domínio frequentemente não podem ser redistribuídos após sua coleta, o que impede a reprodutibilidade de experimentos e limita o avanço colaborativo da área.

Dados sintéticos surgem, então, como uma alternativa para contornar esse problema. Trata-se de dados gerados artificialmente que imitam as propriedades linguísticas, semânticas e estatísticas de dados reais, preservando as características dos dados sem expor diretamente informações originais dos indivíduos. Em domínios sensíveis como a saúde, dados sintéticos têm sido explorados como mecanismo para ampliar conjuntos de dados escassos, balancear classes desiguais e, de forma crescente, viabilizar a distribuição pública de recursos que de outra forma permaneceriam restritos. A premissa central é que, se os dados sintéticos preservarem as propriedades necessárias para treinar modelos de aprendizado de máquina, eles podem substituir funcionalmente os dados originais sem comprometer a privacidade dos usuários que os geraram.

Embora a geração de dados sintéticos tenha recebido atenção crescente nos últimos anos, com aplicações que vão da privatização de texto sob garantias formais (IGAMBERDIEV; HABERNAL, 2023) ao aumento de dados em outros idiomas e domínios (MOHAMMADI; AMIN; TAVAKOLI, 2025; ZHANG et al., 2025), ainda são limitados os estudos que combinam simultaneamente a construção de *corpora* sintéticos em português, a aplicação ao domínio de saúde mental e a motivação central de viabilizar a distribuição pública de dados sensíveis. É precisamente nessa lacuna que se insere o presente trabalho. O *corpus* utilizado, composto por postagens de redes sociais anotadas com sintomas de

depressão, não pode ser disponibilizado publicamente por restrições éticas e de privacidade.

Para isso, propõe-se um pipeline de geração de paráfrases baseado em tradução automática e modelos de linguagem: os textos originais em português são traduzidos para o inglês, submetidos ao modelo PEGASUS (ZHANG et al., 2020a) para geração de paráfrases, retraduzidos ao português e, por fim, selecionados com base em similaridade semântica calculada por meio do BERTScore (ZHANG et al., 2020b). Essa abordagem busca produzir um *corpus* sintético que possa ser disponibilizado publicamente como substituto funcional do *corpus* original.

A hipótese central que orienta a avaliação é que o *dataset* sintético deve ter desempenho similar ao do *dataset* original em tarefas de classificação binária de sintomas de depressão. Se classificadores treinados com dados sintéticos atingirem desempenho equivalente ao de classificadores treinados com dados reais, isso valida a utilidade dos dados gerados como substituto distribuível.

Assim, as principais contribuições deste trabalho são: (i) a proposição de um pipeline para geração de dados sintéticos em português, baseado em tradução automática e parafraseamento; (ii) a construção de uma versão sintética de um *corpus* anotado com sintomas de transtornos de saúde mental; e (iii) a avaliação da capacidade desses dados sintéticos de substituir funcionalmente o *corpus* original em uma tarefa de classificação automática de sintomas de depressão.

Este trabalho está organizado em seis capítulos. O Capítulo 2 apresenta a fundamentação teórica necessária para o desenvolvimento da pesquisa, com ênfase na arquitetura Transformer e nos conceitos de tradução automática neural utilizados na geração de dados sintéticos. O Capítulo 3 revisa e discute os trabalhos relacionados ao tema. O Capítulo 4 descreve o pipeline proposto, detalhando as etapas de tradução automática, geração de paráfrases, seleção de amostras e avaliação por meio de classificadores. O Capítulo 5 apresenta os experimentos realizados, os resultados obtidos e sua análise. Por fim, o Capítulo 6 apresenta as conclusões do trabalho, suas limitações e as perspectivas para trabalhos futuros.

2 Fundamentação Teórica

Este capítulo apresenta os conceitos teóricos que embasam as ferramentas e abordagens utilizadas nesta pesquisa. A organização parte da arquitetura Transformer, base dos modelos empregados (PEGASUS e BERTimbau), e encerra com uma contextualização sobre tradução automática neural e representações distribuídas.

2.1 Conceitos de Privacidade em PLN

No contexto de pesquisas que envolvem dados pessoais, especialmente em domínios como saúde mental, é fundamental compreender os conceitos de dado sensível, desidentificação e reidentificação, que fundamentam algumas das preocupações éticas e legais associadas ao compartilhamento de *corpora*.

2.1.1 Dados Sensíveis

Dados sensíveis são categorias de informação pessoal cujo tratamento inadequado pode causar dano significativo ao indivíduo (Brasil, 2018; European Parliament and Council, 2016). No Brasil, a Lei Geral de Proteção de Dados (LGPD, Lei nº 13.709/2018) (Brasil, 2018) define explicitamente como dados sensíveis aqueles referentes à saúde, origem racial ou étnica, convicção religiosa, opinião política, filiação sindical, dado genético ou biométrico, vida sexual e dados de criança ou adolescente. Postagens em redes sociais que relatam sintomas de depressão, ideação suicida ou tratamento psicológico se enquadram nessa categoria, mesmo quando publicadas originalmente em perfis públicos, pois a publicação voluntária não equivale ao consentimento para uso em pesquisa ou redistribuição (ZIMMER, 2010).

2.1.2 Desidentificação e Reidentificação

A desidentificação é o processo de remover ou modificar informações que permitam identificar diretamente um indivíduo em um conjunto de dados, como nome, endereço, data de nascimento ou instituição (PAGANO et al., 2026). O objetivo é reduzir o risco de que os dados possam ser vinculados à pessoa real que os originou. Em *corpora* textuais, esse processo tipicamente envolve a substituição de entidades nomeadas (pessoas, lugares, instituições) por etiquetas genéricas, como `<NOME>` ou `<CIDADE>`, procedimento adotado, por exemplo, no *corpus* Amive (WILKENS et al., 2026).

No entanto, a desidentificação não garante anonimato completo. A reidentificação é o processo pelo qual informações aparentemente anônimas são cruzadas com outras fontes

para reconstruir a identidade do indivíduo (SWEENEY, 2000). Em textos de redes sociais, mesmo após a remoção de nomes e locais, detalhes contextuais (como referências a eventos específicos, datas, cursos universitários ou combinações incomuns de sintomas) podem ser suficientes para identificar o autor original. Esse risco é amplificado pela disponibilidade crescente de grandes bases de dados públicas que podem ser cruzadas com os dados desidentificados (SWEENEY, 2000).

Mais recentemente, o avanço dos modelos de linguagem de grande porte (LLMs) introduziu uma nova dimensão a esse risco. Diferentemente do cruzamento explícito de bases de dados descrito por Sweeney (2000), LLMs são capazes de inferir atributos pessoais sensíveis (como localização, idade ou ocupação) diretamente a partir de pistas linguísticas sutis presentes no próprio texto, como escolhas vocabulares, referências culturais indiretas ou padrões de escrita, sem necessidade de qualquer fonte externa de cruzamento (STAAB et al., 2024). Patsakis e Lykousas (2023) demonstram ainda que, no contexto pós-LLM, técnicas tradicionais de anonimização textual, antes consideradas suficientes contra adversários humanos, tornam-se vulneráveis a esse tipo de inferência automatizada em larga escala, o que reconfigura o problema da reidentificação de uma tarefa de busca e cruzamento para uma tarefa de inferência estatística sobre o conteúdo textual em si.

É nesse contexto que a geração de dados sintéticos representa uma abordagem complementar à desidentificação: ao substituir os textos originais por paráfrases geradas automaticamente, busca-se produzir um *corpus* que, ainda preservando as propriedades linguísticas e semânticas necessárias para o treinamento de modelos, não corresponda a nenhum texto real publicado por um usuário. Assim, ao substituir integralmente o texto original, é eliminado, em princípio, o risco de reidentificação por cruzamento com fontes externas – embora, como discutido, a proteção oferecida por essa substituição não seja absoluta no contexto atual.

2.2 Arquitetura Transformer

No contexto do processamento automático de texto, a arquitetura Transformer, proposta por Vaswani et al. (2017), representa uma ruptura em relação às abordagens anteriores de processamento sequencial, como as redes recorrentes (RNNs) e as redes de memória de longo prazo (LSTMs). Ao invés de processar *tokens* em sequência, o Transformer processa todos os *tokens* de uma sequência simultaneamente, o que permite maior paralelização e captura de dependências de longo alcance de forma mais eficiente.

Essas características tornam a arquitetura especialmente relevante para tarefas de processamento de linguagem natural que dependem fortemente da compreensão contextual, como tradução automática, geração de texto e análise semântica, que são fundamentais no contexto deste trabalho.

Essa arquitetura é amplamente discutida na literatura de Processamento de Linguagem Natural, sendo detalhada tanto no trabalho original quanto em apresentações didáticas, como a de Paes, Vianna e Rodrigues (2026), que servem como base para a descrição apresentada nesta seção.

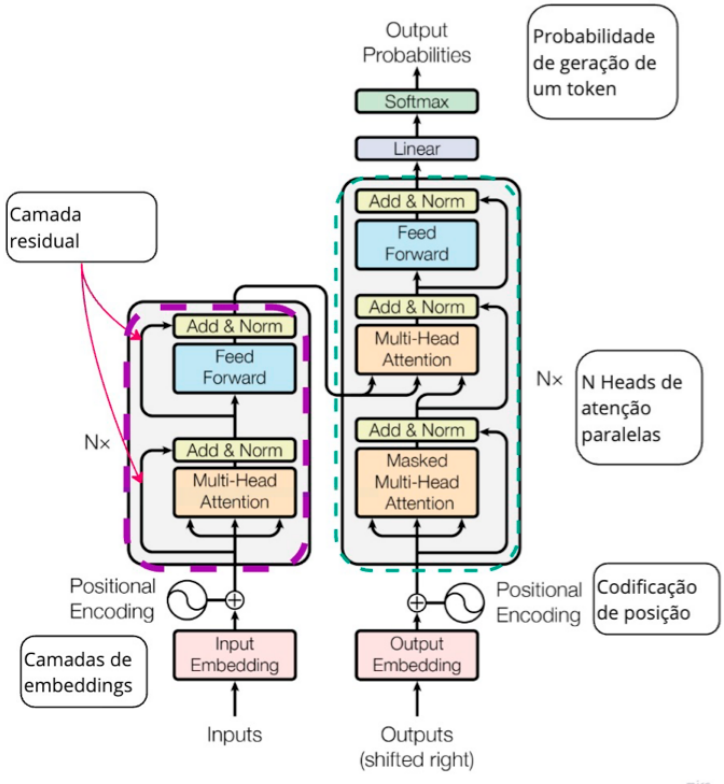
2.2.1 Estrutura Geral

A arquitetura é composta por dois blocos principais: um **encoder**, responsável por transformar a sequência de entrada em representações vetoriais contextualizadas, e um **decoder**, responsável por gerar a sequência de saída com base nessas representações. Modelos voltados à compreensão de texto utilizam apenas o encoder, enquanto modelos voltados à geração utilizam a estrutura encoder-decoder completa.

Cada bloco é constituído por camadas empilhadas, cada uma contendo dois sub-componentes principais: um mecanismo de atenção e uma rede *feed-forward*. A saída de cada sub-componente passa por uma operação de normalização de camada (*layer normalization*) e por uma conexão residual, que soma a entrada ao resultado, estabilizando o treinamento em redes profundas.

A Figura 1 apresenta a arquitetura geral do Transformer.

Figura 1 – Arquitetura Transformer. O pontilhado violeta representa o codificador e o pontilhado verde representa o decodificador.

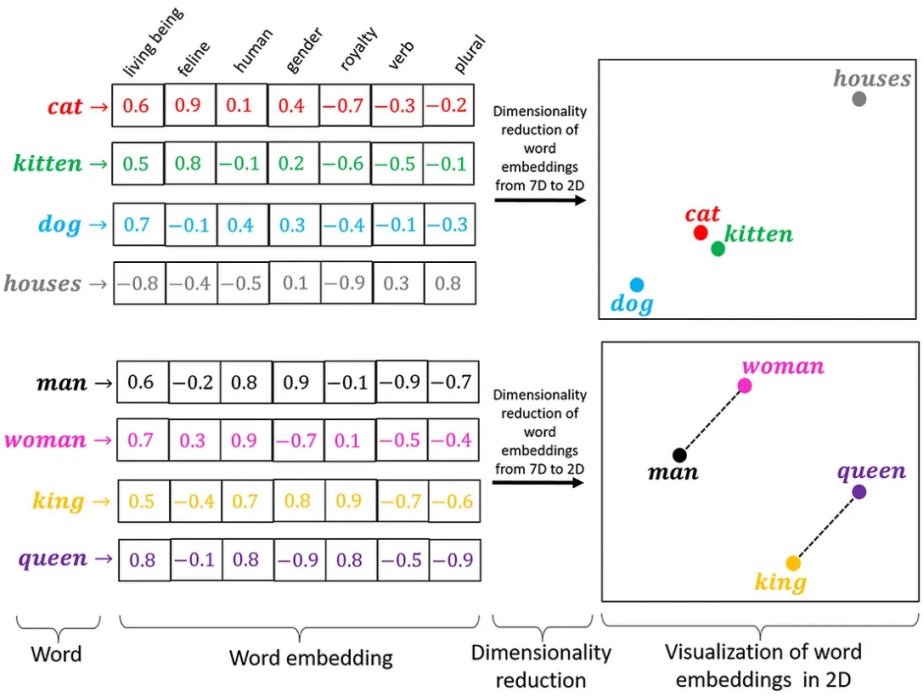


Fonte: (PAES; VIANNA; RODRIGUES, 2026). Os rótulos internos da figura foram mantidos no idioma do original.

2.2.2 Representações Distribuídas

Representações distribuídas, ou *embeddings*, são vetores densos de números reais que codificam o significado de palavras ou sequências de texto em um espaço vetorial contínuo. A ideia central é que itens semanticamente similares devem ocupar regiões próximas nesse espaço, de modo que relações semânticas e sintáticas possam ser capturadas geometricamente. A Figura 2 ilustra esse conceito: palavras com significados próximos, como *cat* e *kitten*, ocupam regiões vizinhas no espaço vetorial, enquanto palavras semanticamente distantes, como *dog* e *houses*, ficam mais afastadas.

Figura 2 – Representações distribuídas: vetores de palavras e sua visualização em espaço bidimensional após redução de dimensionalidade.



Fonte: (MAWATWALA, 2023). Os rótulos internos da figura foram mantidos no idioma do original.

Modelos baseados em Transformer produzem representações **contextualizadas**: ao contrário de abordagens estáticas como Word2Vec (MIKOLOV et al., 2013) ou GloVe (PENNINGTON; SOCHER; MANNING, 2014), em que cada palavra possui um único vetor independente do contexto, os *embeddings* contextualizados variam de acordo com o entorno de cada *token* na sequência. Isso permite que a mesma palavra receba representações distintas conforme seu uso, capturando, por exemplo, ambiguidades semânticas como polissemia.

No contexto desta pesquisa, representações distribuídas contextualizadas são empregadas para calcular a similaridade semântica entre textos, comparando numericamente o quanto o significado de uma paráfrase gerada se aproxima do texto original, o que orienta a seleção das paráfrases mais adequadas para compor o *dataset* sintético.

2.2.3 Codificação Posicional

A entrada do Transformer consiste em sequências de *tokens*, que são primeiro convertidos em vetores densos por meio de uma camada de *embedding*. Como a arquitetura não possui estrutura recorrente, ela não captura naturalmente a ordem dos *tokens*. Para contornar isso, é adicionada uma **codificação posicional** a cada *embedding*, que injeta informação sobre a posição de cada *token* na sequência. Na formulação original, essa codificação utiliza funções senoidais de diferentes frequências:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

onde pos é a posição do *token*, i é a dimensão e d_{model} é a dimensionalidade do modelo. Essa formulação permite que o modelo generalize para sequências mais longas do que as vistas durante o treinamento.

2.2.4 Mecanismo de Atenção

O principal componente da arquitetura Transformer é o mecanismo de atenção, que permite ao modelo atribuir diferentes níveis de importância às palavras de uma sequência ao construir sua representação.

Em vez de processar os *tokens* de forma sequencial, o modelo calcula relações diretas entre todos os elementos da entrada, permitindo que informações relevantes sejam destacadas independentemente da distância entre as palavras no texto.

A formulação utilizada é conhecida como *Scaled Dot-Product Attention*, baseada em três representações vetoriais: consultas (*queries*, Q), chaves (*keys*, K) e valores (*values*, V):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

O produto interno entre Q e K mede a similaridade entre os pares de *tokens*, enquanto o fator de escala $\sqrt{d_k}$ evita que valores muito altos prejudiquem a estabilidade do treinamento. A função *softmax* normaliza os pesos de atenção para que somem 1, produzindo uma distribuição probabilística que determina a contribuição de cada *token* para a representação final. A Figura 3 ilustra esse mecanismo: o *token* “it” presta atenção principalmente a “animal” e “tired”, capturando a relação de correferência na sentença.

2.2.5 Atenção Multi-Cabeça

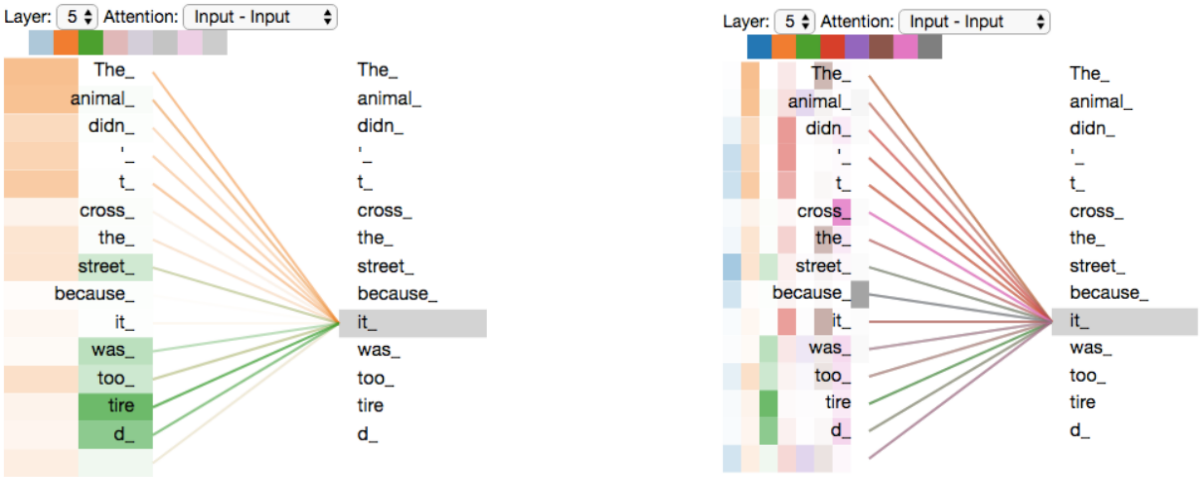
Na prática, o Transformer utiliza atenção multi-cabeça (*multi-head attention*), que executa o mecanismo de atenção em paralelo h vezes, cada vez com projeções lineares distintas de Q, K e V. De forma simples, isso significa que cada cabeça de atenção analisa a mesma entrada de maneira diferente, aplicando diferentes transformações para gerar consultas, chaves e valores.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

$$\text{onde} \quad \text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Com isso, cada cabeça pode capturar diferentes tipos de relações entre os *tokens*, como dependências sintáticas ou associações semânticas, contribuindo para uma representação mais rica da sequência. Isso pode ser observado na Figura 3: as múltiplas cabeças (representadas pelas diferentes cores) distribuem a atenção de formas distintas sobre os mesmos *tokens*, capturando relações complementares entre as palavras.

Figura 3 – Representação do mecanismo de atenção simples (esquerda) e mecanismo de atenção multi-cabeça (direita) para a sentença "The animal didn't cross the street because it was too tired".



Fonte: (BRAINS – Brazilian AI Networks, 2023). Os rótulos internos das figuras foram mantidos no idioma do original.

2.2.6 Atenção no Decoder

No decoder, além da auto-atenção sobre os *tokens* já gerados (com mascaramento para impedir que *tokens* futuros sejam consultados), há um segundo mecanismo denominado atenção cruzada (*cross-attention*), no qual as representações do encoder são incorporadas ao processo de geração. Formalmente, o *cross-attention* segue a mesma formulação do mecanismo de atenção descrito na subseção 2.2.4, com a diferença de que as consultas

(Q) são derivadas dos *tokens* já gerados pelo *decoder*, enquanto as chaves (K) e os valores (V) são derivados das representações produzidas pelo *encoder*. Essa distinção é o que permite ao modelo condicionar cada *token* gerado ao conteúdo específico da sequência de entrada. Na Figura 1, esse mecanismo corresponde à camada "Multi-Head Attention" central do *decoder* – aquela que recebe conexões vindas do *encoder* (representado pelo pontilhado violeta), diferenciando-se da camada de auto-atenção mascarada logo abaixo, cujas conexões são internas ao próprio *decoder*.

Esse mecanismo é essencial em tarefas como tradução automática e geração condicional de texto, pois permite que o modelo utilize a sequência de entrada como referência direta durante o processo de geração. Dessa forma, cada elemento da saída pode ser construído considerando partes específicas da entrada, garantindo maior coerência e alinhamento semântico entre as sequências.

2.3 Tradução Automática Neural

A tradução automática neural (*Neural Machine Translation*, NMT) consiste no uso de redes neurais profundas para traduzir texto de um idioma de origem para um idioma alvo. Ao contrário das abordagens estatísticas anteriores, que decompunham o problema de tradução em componentes separados, os sistemas neurais aprendem a tarefa de forma integrada, mapeando diretamente sequências de origem para sequências alvo a partir de *corpora* paralelos, sem necessidade de recursos linguísticos externos. Assim, têm-se uma rede neural *end-to-end*, sem distinção entre o modelo de tradução e o modelo de língua (CASTILHO; CASELI, 2026).

As palavras são representadas como *embeddings* (vetores densos que codificam propriedades semânticas e sintáticas, descritos em detalhe na subseção 2.2.2) e a geração da tradução é feita sequencialmente, com o modelo considerando tanto a sentença de origem quanto o que já foi produzido na sequência alvo. Com a adoção da arquitetura Transformer, descrita na seção 2.2, os sistemas de tradução neural atingiram o estado da arte, produzindo traduções de alta qualidade, especialmente para pares de idiomas com grande disponibilidade de dados paralelos, como inglês e português (WAY; FORCADA, 2018).

No contexto desta pesquisa, a tradução automática desempenha um papel central no pipeline de geração de paráfrases: os textos originais em português são traduzidos para o inglês, onde modelos de paráfrase (sistemas treinados para reformular textos preservando seu significado original) de maior qualidade estão disponíveis (como o PEGASUS, descrito em detalhe na subseção 4.1.1.2), e os resultados são traduzidos de volta ao português. Essa estratégia de idioma intermediário permite aproveitar recursos desenvolvidos para o inglês para enriquecer um *corpus* originalmente em português.

Assim, a qualidade dos sistemas NMT modernos viabilizou seu uso não apenas para tradução automática, mas também como componente em pipelines de construção e enriquecimento de *corpora*, incluindo tarefas de aumento de dados em PLN. Uma estratégia particularmente relevante nesse contexto é a ***back-translation***: textos em um idioma alvo são traduzidos para um idioma intermediário, processados e retraduzidos ao idioma original, permitindo aproveitar recursos disponíveis no idioma intermediário para enriquecer *corpora* em outros idiomas (WAY; FORCADA, 2018).

No presente trabalho, essa estratégia é adotada no pipeline de geração de paráfrases: os textos originais em português são traduzidos para o inglês, onde um modelo de paráfrase de alta qualidade é aplicado, e os resultados são retraduzidos ao português. A escolha do inglês como idioma intermediário se justifica pela maior disponibilidade de modelos de paráfrase desenvolvidos para esse idioma, permitindo enriquecer um *corpus* originalmente em português, sem abrir mão da qualidade de geração.

2.4 Métricas de Avaliação

Esta seção apresenta as métricas usadas na avaliação automática dos dados sintéticos gerados neste trabalho, seja indiretamente via avaliação extrínseca dos dados por meio de seu uso para treinamento de classificadores, seja diretamente via avaliação intrínseca das propriedades do *dataset* gerado.

2.4.1 Métricas de Classificação

As métricas apresentadas nesta seção são amplamente utilizadas na avaliação de sistemas de classificação e recuperação da informação (MANNING; RAGHAVAN; SCHÜTZE, 2008). Tais métricas são calculadas a partir da matriz de confusão (Figura 4), que relaciona as predições realizadas pelo classificador com os rótulos verdadeiros das instâncias. Em problemas de classificação binária, como é o caso deste trabalho, os resultados podem ser organizados em quatro categorias: verdadeiros positivos (VP), correspondentes às instâncias positivas (1 na Figura 4) corretamente classificadas; verdadeiros negativos (VN), correspondentes às instâncias negativas (0 na Figura 4) corretamente classificadas; falsos positivos (FP), quando uma instância negativa é incorretamente classificada como positiva; e falsos negativos (FN), quando uma instância positiva é incorretamente classificada como negativa.

Figura 4 – Matriz de confusão

	Predito	
Real	**1**	**0**
1	VP	FN
0	FP	VN

A partir desses valores, diferentes métricas podem ser calculadas para avaliar aspectos distintos do desempenho do modelo.

2.4.1.1 Acurácia

A acurácia (*Accuracy*) mede a proporção total de predições corretas realizadas pelo classificador, considerando tanto exemplos positivos quanto negativos (MANNING; RAGHAVAN; SCHÜTZE, 2008). Essa métrica fornece uma visão geral do desempenho do modelo, sendo definida como:

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}$$

Embora seja de fácil interpretação, a acurácia pode ser enganosa em conjuntos de dados desbalanceados, nos quais uma das classes é significativamente mais frequente que a outra.

2.4.1.2 Precisão

A precisão (*Precision*) mede a proporção de instâncias classificadas como positivas que realmente pertencem à classe positiva (MANNING; RAGHAVAN; SCHÜTZE, 2008). Essa métrica é particularmente importante quando falsos positivos possuem elevado custo.

$$\text{Precisão} = \frac{VP}{VP + FP}$$

Valores elevados de precisão indicam que o modelo produz poucas classificações positivas incorretas.

2.4.1.3 Revocação

A revocação (*Recall*), também chamada de sensibilidade, mede a capacidade do modelo de identificar corretamente as instâncias positivas presentes no conjunto de dados (MANNING; RAGHAVAN; SCHÜTZE, 2008).

$$\text{Revocação} = \frac{VP}{VP + FN}$$

Quanto maior a revocação, menor é a quantidade de exemplos positivos que deixam de ser identificados pelo classificador. No contexto deste trabalho, essa métrica é particularmente relevante, uma vez que, em aplicações relacionadas à saúde mental, a não identificação de casos positivos pode ser mais crítica do que a ocorrência de falsos positivos, dado o potencial impacto da omissão de indivíduos em risco.

2.4.1.4 Medida F1

A medida F1 (*F1-score*) combina precisão e revocação em uma única métrica por meio de sua média harmônica, formulação derivada da medida F proposta originalmente por Rijsbergen (1979). Essa medida é particularmente útil quando se deseja equilibrar a capacidade do modelo de evitar falsos positivos e falsos negativos.

$$\text{F1} = 2 \cdot \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}$$

Diferentemente da média aritmética, a média harmônica penaliza situações em que uma das métricas apresenta valor muito inferior à outra, tornando a medida F1 uma alternativa adequada para cenários com classes desbalanceadas.

Neste trabalho, as métricas de acurácia, precisão, revocação e F1 são utilizadas para comparar o desempenho dos classificadores treinados com o *dataset* original e com o *dataset* sintético, permitindo avaliar em que medida os dados gerados preservam a utilidade do *corpus* original para a tarefa de classificação.

2.4.2 Métricas de Propriedades Textuais

As métricas apresentadas até aqui avaliam o desempenho dos classificadores treinados sobre os *datasets* original e sintético. No entanto, para compreender a qualidade intrínseca do *dataset* sintético (independentemente do desempenho em classificação), são utilizadas três métricas adicionais voltadas à avaliação da diversidade lexical das paráfrases geradas: tamanho do vocabulário, D-2 e BLEU (ZHANG et al., 2025).

2.4.2.1 Tamanho do Vocabulário

O tamanho do vocabulário corresponde ao número de tipos lexicais distintos presentes em um conjunto de textos, isto é, o total de palavras únicas após tokenização (ZHANG et al., 2025). Essa métrica fornece uma medida direta da riqueza lexical do *corpus*: um vocabulário maior indica que o conjunto de textos emprega uma variedade maior de termos. No contexto da avaliação de dados sintéticos, o tamanho do vocabulário é calculado separadamente para o *corpus* original e para o *corpus* sintético, permitindo verificar se o processo de paráfrase amplia, mantém ou reduz a diversidade lexical em relação aos textos originais.

2.4.2.2 Diversidade de Bigramas (D-2)

A diversidade de bigramas, denominada D-2, mede a variedade de pares de palavras adjacentes em um conjunto de textos (IPPOLITO et al., 2019). Um bigrama é formado

por duas palavras consecutivas em uma sequência; a métrica é calculada como a proporção de bigramas únicos em relação ao total de bigramas presentes no conjunto:

$$\text{D-2} = \frac{|\text{bigramas únicos}|}{|\text{bigramas totais}|}$$

Valores próximos de 1 indicam alta diversidade, isto é, praticamente todos os pares de palavras são distintos, enquanto valores próximos de 0 indicam que as mesmas combinações se repetem com frequência. Assim como o tamanho do vocabulário, o D-2 é calculado separadamente para o *corpus* original e para o *corpus* sintético, permitindo verificar se o processo de paráfrase introduz variação lexical genuína em relação aos textos originais.

2.4.2.3 BLEU

O BLEU (*Bilingual Evaluation Understudy*) é uma métrica originalmente proposta para avaliação de tradução automática (PAPINENI et al., 2002), que mede a similaridade entre um texto gerado e um texto de referência com base na sobreposição de n-gramas. Para cada n-grama do texto gerado, verifica-se quantas vezes ele aparece no texto de referência; a pontuação final é a média geométrica dessas precisões para diferentes valores de `n`, penalizada quando o texto gerado é mais curto que a referência. A métrica varia entre 0 e 1, sendo que valores mais altos indicam maior similaridade com o texto de referência; valores mais próximos de 0 indicam maior afastamento estrutural e lexical.

No contexto da avaliação de dados sintéticos, o BLEU é utilizado de forma invertida em relação ao seu uso original (ZHANG et al., 2025): em vez de buscar alta similaridade com uma referência desejada, valores mais baixos são preferíveis, pois indicam que as paráfrases geradas se afastam das estruturas lexicais dos textos originais. Cada paráfrase sintética é avaliada contra seu texto original correspondente, e o BLEU médio sobre o conjunto é reportado.

É importante destacar que, ao contrário do tamanho do vocabulário e do D-2, o BLEU não é calculado para o *corpus* original isoladamente, pois a métrica pressupõe a comparação entre dois textos: o gerado (sintético) e o de referência (original).

2.4.2.4 BERTScore

O BERTScore (ZHANG et al., 2020b) é uma métrica de avaliação de geração de texto que, ao contrário do BLEU, não se baseia na sobreposição exata de n-gramas, mas sim na similaridade semântica entre representações contextualizadas obtidas por modelos de linguagem baseados em *Transformer*, como o BERT. Para cada token do texto candidato, é identificado o token mais similar no texto de referência por meio do cálculo de similaridade de cosseno entre seus *embeddings* contextualizados; o processo inverso

também é realizado, identificando para cada token da referência o token mais similar no candidato. A partir desses pareamentos, calculam-se precisão (P_{BERT}) e revocação (R_{BERT}), combinadas na métrica F1:

$$F_{\text{BERT}} = 2 \cdot \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$$

onde P_{BERT} mede a precisão (quanto do texto candidato está semanticamente presente na referência) e R_{BERT} mede a revocação (quanto da referência está coberto pelo candidato). Como essa comparação é feita baseada em representações semânticas em vez de correspondências exatas de palavras, o BERTScore é capaz de identificar similaridade entre um texto e sua sim referência mesmo quando palavras são substituídas por sinônimos ou quando há reformulações estruturais (situações em que o BLEU tenderia a penalizar o texto gerado apesar da preservação de significado). Assim, no BERTScore valores altos são desejáveis, pois indicam preservação de significado entre o texto gerado e sua referência.

2.4.2.5 Distribuição de Categorias Gramaticais

A distribuição de categorias gramaticais é uma medida estrutural utilizada para caracterizar a composição morfossintática de um texto ou conjunto de textos, com base na proporção de ocorrência de cada classe gramatical (como substantivos, verbos, pronomes, adjetivos, advérbios e substantivos próprios, entre outras), de acordo com um conjunto de etiquetas morfossintáticas atribuídas via processo de etiquetação (*part-of-speech tagging*, ou POS *tagging*). Essa medida permite quantificar, para um *corpus* inteiro, qual a proporção de tokens pertencente a cada categoria, possibilitando comparações estruturais entre diferentes *corpora* ou subconjuntos de dados.

No contexto da avaliação de *corpora* sintéticos, a distribuição de categorias gramaticais complementa as métricas de diversidade lexical apresentadas, que avaliam a variação no nível das palavras, oferecendo uma perspectiva sobre a variação gramatical dos textos gerados. Uma alteração acentuada na proporção de determinadas categorias gramaticais entre o *corpus* original e o *corpus* sintético pode indicar mudanças sistemáticas no estilo de escrita introduzidas pelo processo de geração. Uma redução expressiva na proporção de substantivos próprios, por exemplo, pode sinalizar a omissão ou generalização de nomes específicos, enquanto variações nas proporções de verbos e substantivos podem indicar alterações no grau de ação ou descrição predominante nos textos.

2.4.2.6 Índice de Jaccard

O índice de Jaccard, utilizado no contexto deste trabalho para quantificar a diferença lexical entre cada par original-paráfrase, é calculado como a razão entre o tamanho da diferença simétrica e o tamanho da união dos conjuntos de tokens dos dois textos:

$$J_{\text{diff}}(A, B) = \frac{|A \triangle B|}{|A \cup B|} \tag{2.1}$$

onde A e B são os conjuntos de tokens do texto original e da paráfrase, respectivamente, e $A \triangle B$ denota sua diferença simétrica (tokens presentes em apenas um dos dois conjuntos). Por normalizar pelo tamanho do texto original, essa formulação normaliza pela união dos dois vocabulários, tornando a medida simétrica e menos sensível a variações no comprimento do texto. Valores próximos de 1 indicam que praticamente nenhum token é compartilhado entre original e paráfrase, enquanto valores próximos de 0 indicam alta sobreposição lexical.

3 Trabalhos Relacionados

Este capítulo apresenta os trabalhos que fundamentam e contextualizam a presente pesquisa. A organização parte da motivação clínica e social do problema, passando pela questão da privacidade que torna o *corpus* original intransferível, e chega-se às estratégias técnicas de geração e avaliação de dados sintéticos adotadas na literatura e no presente trabalho.

3.1 PLN Aplicado à Saúde Mental

Transtornos mentais como depressão e ansiedade figuram entre as principais causas de incapacidade no mundo, e a identificação precoce de seus sinais representa um desafio clínico relevante. Com a popularização das redes sociais, o conteúdo publicado por usuários passou a ser explorado como fonte de dados para tarefas de triagem e monitoramento em saúde mental. Isso representa uma vertente crescente em PLN que busca detectar automaticamente sinais e sintomas de transtornos de saúde mental em texto.

Paraboni e Caseli (2024) apresentam um panorama dessa área no âmbito da detecção computacional de transtornos de saúde mental a partir de redes sociais, com foco no português brasileiro. Os autores situam o problema em um contexto de crescente urgência: segundo dados do Global Burden of Disease Study de 2019, a depressão é o segundo fator que mais impacta na queda da expectativa de vida da população global, e a OMS estima que apenas um quarto dos indivíduos com transtornos mentais recebem atendimento adequado. No Brasil, a prevalência da depressão cresceu 36,7% entre 2013 e 2019, e durante a pandemia de Covid-19 estudos registraram sintomas depressivos em mais de 70% dos adultos avaliados. Ao mesmo tempo, indivíduos com transtornos de saúde mental são usuários regulares de redes sociais em proporção similar à população geral, e frequentemente recorrem a essas plataformas em busca de suporte, o que torna o conteúdo publicado nelas uma fonte relevante para triagem e monitoramento computacional. Os autores distinguem duas granularidades de análise: a classificação em nível autoral, que examina a coleção de postagens de um indivíduo para estimar seu estado de saúde mental, e a classificação em nível textual, que examina postagens individuais para detectar sintomas específicos. Destacam ainda que *corpora* nesse domínio geralmente não são distribuíveis publicamente, seja por restrições éticas ou pelos termos de uso das plataformas, obstáculo que motiva diretamente o presente trabalho: a geração de dados sintéticos como mecanismo para viabilizar a distribuição de um *corpus* funcional sem expor os textos originais dos usuários.

Alhamed et al. (2024) aprofundam essa questão ao propor um esquema de anotação

estruturado para classificar níveis de severidade de depressão e sintomas específicos em postagens de usuários. O trabalho define diretrizes detalhadas para anotadores humanos identificarem sinais clínicos em conteúdo gerado em redes sociais, produzindo um *corpus* anotado com categorias de severidade. Os resultados demonstram boa consistência entre anotadores e a viabilidade de construir recursos clinicamente válidos a partir de texto não estruturado. Esse trabalho é relevante para a presente pesquisa por evidenciar que as categorias clínicas expressas nos textos são sutis e precisam ser preservadas durante qualquer processo de geração sintética. Uma paráfrase que altere o tom ou a intensidade de uma postagem pode modificar seu rótulo de forma indesejada, comprometendo a utilidade do dado gerado.

3.2 Privacidade e a Necessidade de Dados Sintéticos

Um obstáculo recorrente em pesquisas de PLN voltadas à saúde mental é a dificuldade de acesso e distribuição de dados. Postagens em redes sociais, mesmo sendo tecnicamente públicas, carregam informações sensíveis sobre saúde mental de indivíduos identificáveis, o que levanta preocupações éticas e legais quanto ao seu compartilhamento irrestrito. Na prática, *corpora* construídos nesse domínio frequentemente não podem ser redistribuídos após sua coleta, o que impede a reprodutibilidade de experimentos e limita o avanço colaborativo da área.

Este é precisamente o contexto do *corpus* utilizado nesta pesquisa: os dados originais, coletados de redes sociais e anotados com categorias de saúde mental, já passaram por processo de anonimização, mas não podem ser disponibilizados publicamente por restrições éticas e de privacidade. Nesse contexto, a geração de dados sintéticos vai além do objetivo principal de aumentar de dados, mas garante distribuição segura dos dados existentes no que diz respeito à privacidade dos usuários que os originaram.

Igamberdiev e Habernal (2023) abordam esse problema de forma mais formal com o modelo DP-BART, que propõe a reescrita de textos sob a garantia matemática de Privacidade Diferencial Local (LDP). O modelo, baseado na arquitetura BART, reescreve cada texto de modo a preservar o significado geral enquanto mascara informações potencialmente identificáveis. O grau de proteção é controlado pelo parâmetro ε: valores menores garantem maior privacidade, mas introduzem maior distorção semântica. Os experimentos mostram que variantes com poda de parâmetros (DP-BART-PR e PR+) oferecem melhor equilíbrio entre privacidade e qualidade textual, embora valores muito baixos de ε ainda comprometam o desempenho em tarefas *downstream* (tarefas de classificação para as quais os dados reescritos serão utilizados, como a detecção propriamente dita de transtornos de saúde mental).

Embora o presente trabalho não implemente privacidade diferencial, o DP-BART

representa uma direção relevante para trabalhos futuros. Os dados originais aqui utilizados passaram previamente por um processo de anonimização com remoção de referências a lugares, instituições, eventos, datas e cursos (explicado em mais detalhes na seção 4.2) que, no entanto, não oferece garantias formais contra reidentificação. Assim, a combinação da geração sintética proposta neste trabalho com técnicas de privacidade diferencial como o DP-BART poderia oferecer uma camada adicional de proteção, produzindo um *corpus* com garantias matematicamente verificáveis contra reidentificação.

3.3 Geração de Dados Sintéticos em Texto

Diversas estratégias foram exploradas na literatura para gerar dados sintéticos que ampliem *corpora* escassos preservando a distribuição das classes originais.

Mohammadi, Amin e Tavakoli (2025) propõem o uso de Redes Generativas Adversariais (GANs) para ampliar um *dataset* de análise de sentimentos em persa, língua com poucos recursos anotados. A abordagem combina um gerador baseado em Transformers com um discriminador que utiliza *embeddings* BERT para avaliar o realismo dos textos produzidos. Técnicas como inserção de sinônimos e substituição de pronomes são aplicadas na geração, e o processo iterativo do discriminador força o gerador a produzir textos progressivamente mais plausíveis. Os resultados mostram ganhos de desempenho nos classificadores treinados com os dados aumentados, especialmente nas classes com menos exemplos. Uma limitação relevante é a dependência de recursos lexicais específicos do idioma, como dicionários de sinônimos, o que dificulta a transferência direta do método para outros contextos linguísticos, incluindo o português brasileiro.

Enquanto a abordagem de Mohammadi, Amin e Tavakoli (2025) prioriza o volume de dados gerados, Zhang et al. (2025) abordam um problema central em qualquer tarefa de geração de dados sintéticos: a produção de amostras ambíguas, ou seja, que poderiam pertencer a mais de uma classe. Em tarefas de detecção fina de intenções, onde classes semanticamente próximas são facilmente confundidas, os autores propõem um framework iterativo de geração diferencial com feedback contrastivo. A ideia central é que, em vez de gerar exemplos para cada classe de forma independente, o modelo é instruído a considerar explicitamente as diferenças entre classes potencialmente confusas antes de produzir novos exemplos, reduzindo a sobreposição semântica entre elas. O processo incorpora refinamento por rubrica, isto é, um conjunto de critérios explícitos que avaliam validade (se o exemplo realmente corresponde à classe esperada) e diversidade (se os exemplos gerados são suficientemente variados entre si), e ciclos iterativos de feedback entre um modelo grande (LLM), responsável pela geração, e um modelo menor (SLM), que sinaliza quais exemplos ainda causam confusão. Os resultados superam abordagens de geração direta em configurações *zero-shot*, *few-shot* e *full-shot* em três *datasets*, com ganhos especialmente

expressivos em classes com alta sobreposição semântica. A preocupação com a qualidade semântica dos dados gerados, e não apenas com seu volume, é diretamente relevante para o presente trabalho: no domínio de saúde mental, a distinção entre classes como presença e ausência de sintomas depressivos pode ser igualmente sutil, e dados sintéticos que borrem essa fronteira comprometem diretamente a utilidade dos classificadores treinados com eles.

Se os dois trabalhos anteriores tratam majoritariamente do inglês e do persa, Johansson et al. (2026) deslocam a questão para o português, apresentando o Meta4XNLI-ptBR, o primeiro *corpus* aberto com anotação humana para detecção de metáforas em nível de *token* no português brasileiro. O trabalho não tem como objetivo o aumento de dados para classificação, mas é diretamente relevante por demonstrar a viabilidade de construir *corpora* em português via tradução automática e por propor um critério de seleção baseado em métrica de qualidade, estrutura análoga à do presente trabalho. O pipeline combina tradução automática de instâncias do *corpus* Meta4XNLI via LLMs (Llama 4 Maverick, Llama 4 Scout e Sabiá 3.1) com anotação humana posterior. Para cada instância, a tradução com maior pontuação na métrica XCOMET-XL é selecionada entre os candidatos gerados pelos diferentes modelos. Os resultados indicam que modelos treinados especificamente em português (Sabiá 3.1) apresentam vantagem sutil na tradução de expressões metafóricas. Uma limitação importante é que o *corpus* resultante cobre apenas 13,39% do Meta4XNLI original, e a tradução automática introduz disfluências especialmente em contextos figurativos (metáforas). A abordagem é especialmente relevante para o presente trabalho pela analogia estrutural entre o uso do XCOMET-XL como critério de seleção da melhor tradução e o uso do BERTScore como critério de seleção da melhor paráfrase. Em ambos os casos, uma métrica de similaridade semântica orienta a escolha entre candidatos gerados automaticamente.

Ainda no contexto do português, Pellicer et al. (2022) propõem o PTT5-Paraphraser, um modelo baseado na arquitetura T5 pré-treinado em português (PTT5) e ajustado para a tarefa de paráfrase sobre o *corpus* TaPaCo, composto por aproximadamente 36.000 pares de paráfrases em português. O modelo busca equilibrar dois objetivos frequentemente conflitantes na geração de paráfrases: fidelidade semântica ao texto original e diversidade lexical das paráfrases geradas. Os autores conduzem avaliações automáticas e humanas da qualidade das paráfrases produzidas, além de um experimento de aumento de dados que demonstra ganhos de desempenho em classificadores treinados com os dados aumentados. O trabalho é diretamente relevante para o presente estudo por propor uma abordagem de paráfrase nativa em português, sem dependência de tradução automática como idioma intermediário. No entanto, o PTT5-Paraphraser foi ajustado sobre um *corpus* de menor escala e de domínio geral em comparação aos dados utilizados no pré-treinamento do PEGASUS, o que pode resultar em menor capacidade de reformulação sintática e diversidade lexical, aspectos centrais para a geração de um *corpus* sintético funcionalmente útil. Assim, no contexto do trabalho, ele é investigado como possível direção para trabalhos futuros

(seção 6.2).

Apesar de empregarem técnicas distintas, os quatro trabalhos compartilham uma mesma tensão central na geração de dados sintéticos: produzir exemplos suficientemente diversos sem comprometer sua validade em relação à classe ou ao significado original. Cada um resolve essa tensão por meio de um mecanismo distinto de seleção automática entre candidatos gerados: Mohammadi, Amin e Tavakoli (2025) recorrem ao discriminador da própria GAN, que filtra exemplos pela plausibilidade frente aos dados reais; Zhang et al. (2025) utilizam um processo iterativo de refinamento por rubrica e feedback contrastivo entre modelos de diferentes portes, que descarta exemplos ambíguos entre classes; Pellicer et al. (2022) avaliam a qualidade das paráfrases geradas por métricas automáticas e principalmente avaliação humana; e Johansson et al. (2026) selecionam, entre múltiplas traduções candidatas, aquela com maior pontuação na métrica XCOMET-XL. Essa última abordagem é estruturalmente a mais próxima do pipeline proposto no presente trabalho, em que a seleção da melhor paráfrase candidata também é feita por meio de uma métrica de similaridade semântica automática (BERTScore), reforçando que o problema de balancear diversidade e fidelidade na geração sintética atravessa diferentes idiomas, domínios e técnicas de geração.

3.4 Avaliação de Dados Sintéticos

Verificar se os dados gerados sinteticamente são efetivamente úteis para as tarefas às quais se destinam constitui uma etapa fundamental do processo de geração, antecedendo sua utilização ou disponibilização final.

Chim, Ive e Liakata (2025) propõe um quadro unificado de avaliação para dados gerados a partir de textos de usuários, estruturado em duas dimensões. A avaliação intrínseca examina os textos em si, considerando preservação de estilo, preservação de significado e divergência entre os exemplos gerados. A avaliação extrínseca mede o impacto dos dados sintéticos em tarefas concretas de PLN: modelos são treinados nos dados gerados e seu desempenho em tarefas *downstream* (como classificação de texto) é comparado ao de modelos treinados com dados reais, verificando se os dados sintéticos preservam a utilidade do *corpus* original. O quadro considera também o risco de privacidade, avaliando em que medida os textos gerados poderiam ser utilizados para reidentificar os indivíduos que produziram os dados originais. Os resultados indicam que a avaliação extrínseca, especialmente o desempenho em tarefas *downstream*, é o critério mais informativo para determinar a utilidade prática dos dados. Uma limitação é que o quadro foi desenvolvido e validado em inglês, exigindo adaptação para outros idiomas.

Zhang et al. (2025) complementam essa visão aplicando as métricas de tamanho do vocabulário, D-2, BLEU invertido (descritos na subseção 2.4.2) e fidelidade para avaliar a

qualidade dos dados sintéticos gerados por seu *framework*. Os resultados mostram que o método proposto produz dados com maior vocabulário e D-2 e menor BLEU do que os *baselines*, indicando maior diversidade lexical, ao mesmo tempo em que mantém fidelidade comparável. Essas mesmas métricas de diversidade são adotadas no presente trabalho para caracterizar o *corpus* sintético gerado, permitindo verificar se o pipeline proposto introduz variação lexical genuína em relação aos textos originais.

A estratégia principal de avaliação adotada nesta pesquisa é, no entanto, extrínseca, alinhada às recomendações de Chim, Ive e Liakata (2025): tanto o *dataset* original quanto o *dataset* sintético são utilizados para treinar classificadores binários independentes, e os resultados são comparados com as mesmas métricas. A hipótese central é que o *dataset* sintético deve reproduzir (ou superar) o desempenho do original em termos de acurácia, precisão, revocação e F1, validando sua utilidade como substituto distribuível do *corpus* que não pode ser compartilhado.

3.5 Síntese

Tabela 1 – Comparativo dos trabalhos relacionados

Trabalho	**Idioma**	**Domínio**	**Geração**	**Avaliação**
Paraboni e Caseli (2024)	PT-BR	Saúde mental	—	Downstream
Alhamed et al. (2024)	Inglês	Saúde mental	—	IAA
Igamberdiev e Habernal (2023)	Inglês	Geral	DP-BART	Downstream
Mohammadi et al. (2025)	Persa	Sentimentos	GAN	Downstream
Zhang et al. (2025)	Inglês	Geral	Diferencial	D-2, BLEU, F1
Pellicer et al. (2022)	PT-BR	Geral	Paráfrase (PTT5)	Automática + Humana
Johansson et al. (2026)	PT-BR	Geral	Tradução + LLM	IAA, XCOMET-XL
Chim, Ive e Liakata (2025)	Inglês	Geral	Paráfrase, LM	Intrínseca + Downstream

Os trabalhos apresentados convergem para os desafios centrais desta pesquisa: a escassez de dados anotados em português para saúde mental, as restrições de privacidade que impedem sua redistribuição, e a necessidade de estratégias de geração e avaliação que garantam a validade clínica e linguística dos dados produzidos. Assim, a escolha por combinar tradução automática, paráfrase via PEGASUS e seleção por similaridade semântica, em vez de estratégias alternativas, decorre de duas considerações práticas.

Primeiramente, modelos de paráfrase nativos para o português, embora existentes (a exemplo do PTT5-Paraphraser (PELLICER et al., 2022)), são tipicamente treinados em *corpora* de menor escala e diversidade do que os disponíveis para o inglês, o que motivou a adoção do inglês como idioma intermediário para acessar modelos de paráfrase mais robustos, como o PEGASUS. Em segundo lugar, optou-se por incluir explicitamente uma etapa de paráfrase, em vez de empregar apenas *back-translation* simples (tradução seguida de retradução, sem reformulação adicional), pois sistemas de tradução automática neural modernos tendem a produzir traduções de alta fidelidade e baixa variação, o que resultaria

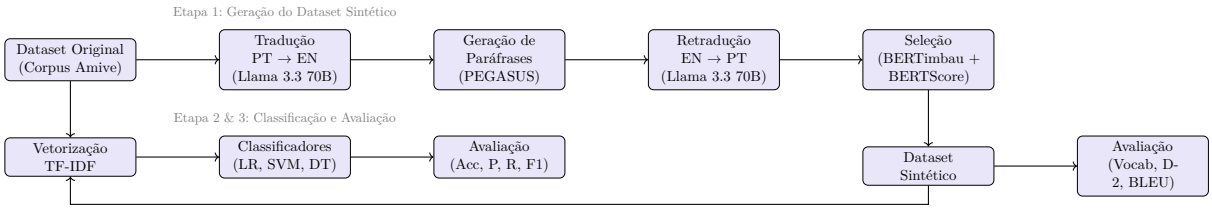
em textos retraduzidos muito próximos aos originais e, portanto, insuficientemente diversos para os fins de privacidade e variação textual buscados neste trabalho (discutido em mais detalhe na seção 6.2). Com isso, o *pipeline* proposto busca equilibrar a robustez dos modelos disponíveis em inglês com a necessidade de diversidade textual genuína, produzindo um *corpus* sintético que possa ser disponibilizado publicamente como substituto funcional do *corpus* original.

4 *Pipeline* para geração de dados sintéticos

Este capítulo descreve as etapas propostas para a geração de dados sintéticos no domínio da saúde mental. Para isso, são descritos os métodos e os materiais utilizados.

O *pipeline* proposto é composto por três etapas sequenciais: geração do *dataset* sintético, treinamento dos classificadores nos *datasets* original e sintético e avaliação comparativa entre seus resultados. A Figura 5 ilustra esse fluxo.

Figura 5 – *Pipeline* experimental: geração do *dataset* sintético e avaliação por classificação binária.



4.1 Métodos

Esta seção descreve os métodos empregados em cada etapa da pesquisa: a geração do *dataset* sintético e a avaliação de sua utilidade por meio de classificação supervisionada.

4.1.1 Geração do *Dataset* Sintético (Etapa 1)

O *pipeline* proposto neste trabalho para a geração de paráfrases é composto por três etapas principais: tradução para o inglês, geração de paráfrases e retradução para o português, seguidas de uma etapa de seleção da melhor paráfrase candidata.

4.1.1.1 Tradução Automática via LLM

A primeira etapa consiste na tradução dos textos originais em português para o inglês. Essa escolha se justifica pela maior disponibilidade de modelos de paráfrase de alta qualidade em inglês, em particular o PEGASUS (detalhado na subseção 4.1.1.2). Assim, ao passar pelo inglês como idioma intermediário, é possível aproveitar um modelo mais robusto do que os atualmente disponíveis para o português, retraduzindo o texto parafraseado e mantendo tanto a entrada quanto a saída do processo no idioma original.

A tradução é realizada por meio do modelo Llama 3.3 70B (AI@Meta, 2024), acessado via plataforma Groq[1], com processamento em lote: múltiplos textos são enviados

[1] <https://groq.com>

em uma única requisição, numerados sequencialmente, e as traduções são extraídas da resposta por meio de expressões regulares. Ao final do *pipeline*, as paráfrases geradas em inglês são traduzidas de volta ao português pelo mesmo mecanismo. O fluxo de tal processo é representado pela Figura 13.

É importante pontuar que a qualidade da tradução automática não foi avaliada de forma isolada. Essa decisão decorre de uma limitação metodológica inerente ao *pipeline*: avaliar a tradução PT→EN com o BERTScore exigiria um modelo multilíngue como referência, em vez do BERTimbau (especializado em português), o que tornaria a comparação assimétrica em relação ao critério de seleção final. Métricas como o COMET, por sua vez, pressupõem um texto de referência humano para o par de idiomas avaliado, recurso não disponível neste contexto. Uma alternativa metodologicamente viável seria gerar múltiplas traduções por instância e selecionar a de maior qualidade antes do parafraseamento (estratégia análoga à adotada por Johansson et al. (2026) com o XCOMET-XL) ou avaliar o *pipeline* de *back-translation* puro (sem paráfrase), no qual a qualidade da tradução poderia ser inferida pela similaridade entre o texto original e o retraduzido. Tais direções constituem extensões naturais deste trabalho, discutidas na seção 6.2.

4.1.1.2 Geração de Paráfrases com PEGASUS

O PEGASUS (*Pre-training with Extracted Gap-sentences for Abstractive SUmmarization*), proposto por Zhang et al. (2020a), é um modelo baseado na arquitetura encoder-decoder do Transformer, pré-treinado para tarefas de geração de texto. Sua estratégia de pré-treinamento, denominada *Gap Sentence Generation* (GSG), consiste em remover sentenças completas de um documento e treinar o modelo para reconstruí-las a partir do restante do texto, exigindo compreensão global do conteúdo. O PEGASUS foi originalmente desenvolvido para tarefas de sumarização abstrativa, isto é, produção de resumos de textos que não se limite a copiar trechos do original, mas sim reescrever o conteúdo com novas palavras. No entanto, demonstrou forte capacidade de paráfrase, pois ambas as tarefas exigem reformulação da expressão aliada a preservação do significado central.

Para cada texto traduzido ao inglês, o PEGASUS gera múltiplas paráfrases candidatas por meio de *diverse beam search* (VIJAYAKUMAR et al., 2016). A busca em feixe (*beam search*) é uma estratégia de geração de texto que, em vez de escolher sempre a palavra mais provável a cada passo, mantém as N sequências mais promissoras em paralelo (chamadas de feixes) e as expande simultaneamente, descartando as menos prováveis ao final de cada passo. Na variante *diverse beam search*, os feixes são divididos em grupos e uma penalidade de diversidade é aplicada entre grupos, desincentivando que feixes do mesmo grupo gerem sequências semelhantes entre si. Isso resulta em paráfrases variadas tanto em vocabulário quanto em estrutura sintática. O fluxo de tal processo é representado

pela Figura 14.

4.1.1.3 Seleção com BERTimbau

Após a retradução ao português das paráfrases candidatas restantes (subseção 4.1.1.1), a seleção da mais adequada é feita com base no BERTScore Dentre as paráfrases candidatas restantes, a seleção da mais adequada é feita com base no BERTScore (subseção 2.4.2.4), calculado com o BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020) como modelo de representação. O BERTimbau é uma versão do BERT pré-treinada especificamente em português brasileiro, com aproximadamente 2,7 bilhões de tokens provenientes de fontes diversas, incluindo a Wikipedia em português e o *corpus* BrWaC. Por ser especializado no idioma, o BERTimbau captura padrões morfossintáticos e semânticos do português brasileiro com maior fidelidade do que modelos multilíngues genéricos, o que o torna adequado para avaliar a preservação de significado entre as paráfrases candidatas e os textos originais.

A paráfrase com maior F_{BERT} em relação ao texto original, calculado conforme descrito na subseção 2.4.2.4, é selecionada como a paráfrase final daquela instância. O fluxo completo de tal processo é representado pela Figura 15.

4.1.2 Classificação e Avaliação (Etapas 2 e 3)

4.1.2.1 TF-IDF para Representação dos Textos

Os textos são representados por meio de TF-IDF (*Term Frequency–Inverse Document Frequency*), uma medida clássica de PLN que pondera a frequência de cada termo em um documento pela raridade desse termo na coleção como um todo. Formalmente, o peso de um termo t em um documento d é dado por:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log \frac{N}{\text{DF}(t)}$$

onde $\text{TF}(t, d)$ é a frequência do termo no documento, N é o número total de documentos e $\text{DF}(t)$ é o número de documentos que contêm o termo. Termos muito frequentes em toda a coleção (como artigos e preposições) recebem pesos baixos, enquanto termos discriminativos e específicos de poucos documentos recebem pesos altos.

4.1.2.2 Classificadores

Quatro classificadores são treinados sobre as representações TF-IDF:

1. **Regressão Logística** é um modelo linear que estima a probabilidade de uma instância pertencer à classe positiva por meio de uma função sigmoide aplicada a

uma combinação linear dos atributos. Apesar de sua simplicidade, é amplamente utilizada como baseline forte em tarefas de classificação de texto.

2. **SVM** (*Support Vector Machine*) com kernel linear busca o hiperplano de separação que maximiza a margem entre as classes. Em espaços de alta dimensão (como os produzidos por TF-IDF) o SVM linear é frequentemente competitivo com modelos mais complexos.

3. **Naive Bayes** é um modelo probabilístico que estima a classe de uma instância com base na distribuição dos atributos, assumindo independência entre eles. Apesar de essa suposição ser simplificada e raramente válida em linguagem natural, o modelo costuma apresentar bom desempenho em tarefas de classificação de texto, especialmente com representações esparsas como TF-IDF. Sua inclusão é importante por ser um *baseline* sensível a mudanças no vocabulário, ajudando a avaliar se padrões estatísticos básicos são preservados entre os conjuntos de dados original e sintético.

4. **Random Forest** é um modelo de conjunto (*ensemble*) que combina múltiplas árvores de decisão treinadas sobre subamostras aleatórias dos dados e dos atributos, agregando suas predições por votação majoritária. Essa estratégia reduz a variância em relação a uma árvore isolada (que tende ao *overfitting* e é instável diante de pequenas variações nos dados de entrada), produzindo um modelo mais robusto e confiável.

   A robustez é especialmente relevante no contexto de uso de representações TF-IDF: como os processos de parafraseamento e retradução podem introduzir sinônimos e reformulações que alteram as frequências dos termos, uma árvore isolada poderia apresentar divisões completamente diferentes entre o *dataset* original e o sintético, tornando a comparação entre os dois cenários pouco confiável. O Random Forest suaviza esse efeito ao distribuir as decisões entre múltiplas árvores, capturando padrões lexicais globais que se mantêm mesmo diante de variação vocabular.

A escolha conjunta desses quatro modelos visa cobrir diferentes hipóteses de aprendizado: linear probabilística (Regressão Logística), margem máxima (SVM), probabilística ingênua baseada em independência condicional (Naive Bayes) e particionamento hierárquico (Random Forest), permitindo avaliar de forma mais robusta o impacto dos dados sintéticos no desempenho da classificação.

Vale reforçar que tal escolha não teve como objetivo identificar a configuração (algoritmo, parâmetros etc.) com melhor desempenho para a tarefa, mas sim fornecer diferentes perspectivas para a avaliação dos conjuntos de dados. Como o foco deste trabalho está na comparação entre o *dataset* original e o *dataset* sintético, buscou-se empregar

modelos amplamente utilizados e com características distintas de aprendizado, de modo a verificar se os efeitos observados se mantêm em diferentes abordagens de classificação. Dessa forma, não foi realizada uma busca exaustiva por algoritmos ou configurações ótimas, uma vez que o interesse principal está em avaliar a capacidade do conjunto sintético de reproduzir o comportamento do conjunto original em uma tarefa de classificação binária. Salvo exceções explicitamente justificadas na subseção 5.2.2, os classificadores foram utilizados com os parâmetros padrão fornecidos pela biblioteca `Scikit-learn`, escolha consistente com o objetivo de comparação entre *datasets*, e não de otimização de desempenho absoluto.

4.1.2.3 Protocolo de Avaliação Extrínseca

O protocolo de avaliação tem como objetivo comparar diretamente o desempenho obtido com o *dataset* original e com o *dataset* sintético. Ambos os *datasets* são divididos nos mesmos conjuntos de treino e teste: a divisão é feita por identificador de instância no *dataset* original e replicada no sintético, garantindo que textos da mesma postagem não apareçam simultaneamente em treino e teste, e que a comparação entre todos os cenários seja feita sobre os mesmos exemplos. São avaliados quatro cenários, descritos a seguir.

1. Cenário original: o classificador é treinado e testado sobre o *dataset* original.
2. Cenário sintético: o classificador é treinado e testado sobre o *dataset* sintético, utilizando os mesmos identificadores de treino e teste do cenário original.
3. Cenário cruzado sintético-original: o classificador é treinado no conjunto de treino do *dataset* sintético e testado no conjunto de teste do *dataset* original (os mesmos identificadores de instância de treino e teste definidos no cenário original), permitindo avaliar a capacidade de generalização dos dados sintéticos para exemplos reais. Este é o cenário mais relevante para validar a utilidade do *corpus* sintético como substituto distribuível do *corpus* original.
4. Cenário cruzado original-sintético: o classificador é treinado no conjunto de treino do *dataset* original e testado no conjunto de teste do *dataset* sintético (novamente os mesmos identificadores de instância), permitindo avaliar, de forma complementar, se um classificador treinado em dados reais consegue generalizar para o vocabulário e a estrutura introduzidos pelas paráfrases.

O desempenho é medido por acurácia, precisão, revocação e F1, e os resultados obtidos nos quatro cenários são comparados entre si: se o classificador treinado com dados sintéticos atingir desempenho equivalente ao treinado com dados originais, isso indica que as paráfrases geradas preservam as propriedades linguísticas e semânticas necessárias para a tarefa, validando sua utilidade como substituto distribuível do *corpus* original.

4.1.2.4 Protocolo de Avaliação Intrínseca

Além da avaliação extrínseca por classificação, o *dataset* sintético é adicionalmente avaliado intrinsecamente por meio de três métricas de diversidade lexical, descritas na subseção 2.4.2: tamanho do vocabulário, D-2 e BLEU invertido. O tamanho do vocabulário e o D-2 são calculados separadamente para o *corpus* original e para o *corpus* sintético e comparados entre si, permitindo verificar se o processo de paráfrase introduz variação lexical genuína em relação aos textos originais. O BLEU é calculado para cada par original-paráfrase individualmente e reportado como média sobre o conjunto, medindo o afastamento estrutural de cada paráfrase em relação ao seu texto de origem. Juntas, essas métricas permitem caracterizar a diversidade do *corpus* sintético tanto ao nível do conjunto, isto é, verificando se as paráfrases são variadas entre si, quanto ao nível de cada instância, verificando se cada paráfrase reformulou genuinamente o texto original.

4.2 Materiais

4.2.1 *Corpus* Amive

O *corpus* utilizado nesta pesquisa é o derivado do projeto Amive (*Amigo Virtual Especializado*), desenvolvido na Universidade Federal de São Carlos (UFSCar) entre 2021 e 2023 por uma equipe multidisciplinar de pesquisadores das áreas de computação, psicologia, terapia ocupacional e psiquiatria (PARABONI; CASELI, 2024). O *corpus* é composto por postagens anônimas coletadas de páginas públicas de segredos universitários no Facebook, plataforma em que estudantes universitários brasileiros publicam relatos pessoais de forma anônima por meio de formulários. A coleta foi realizada via plataforma Crowdtangle, abrangendo postagens entre janeiro de 2012 e dezembro de 2021, utilizando palavras-chave associadas a depressão e ideação suicida: “suicídio”, “depressão”, “me corto”, “vontade de viver”, “me matar” e “quero morrer”.

Das postagens coletadas, 780 foram selecionadas para anotação manual. Dado o conteúdo sensível, antes da anotação foi realizado um processo adicional de anonimização manual, substituindo referências a lugares, instituições, eventos, datas e cursos por etiquetas genéricas como `<CIDADE>`. As postagens foram anotadas por quatro estudantes de psicologia, psiquiatria ou terapia ocupacional, supervisionados por especialistas em saúde mental.

A anotação foi realizada em dois níveis. No nível autoral, cada postagem foi classificada como indicativa de um Possível Perfil Depressivo (PPD) ou não. No nível textual, trechos das postagens foram anotados com um ou mais dos 18 sintomas de depressão definidos pela equipe clínica do projeto, além de 3 sinais adicionais (fatores de risco, fatores protetivos e morte ou suicídio de outro). O *corpus* foi utilizado por Mendes e Caseli (2024) em experimentos de classificação automática dos sintomas de depressão

nele anotados, com técnicas que vão de modelos clássicos de aprendizado de máquina a modelos baseados em Transformers, demonstrando a viabilidade do *corpus* para tarefas de classificação em nível textual. Para o presente trabalho, é utilizada a versão revisada do *corpus* descrita em (WILKENS et al., 2026), que padronizou as fronteiras dos trechos anotados e verificou a presença de todos os segmentos no *corpus* final. Essa versão contém 604 postagens, sendo 336 classificadas como PPD e 268 como não-PPD.

A Tabela 2 apresenta a distribuição de instâncias anotadas por sintoma. Os sintomas com maior número de instâncias são Tristeza ou humor depressivo (469), Desamparo, prejuízo social ou solidão (377) e Suicídio ou auto-extermínio (278), enquanto os menos frequentes são Agitação ou inquietação (15) e Déficit de atenção ou memória (16). Essa distribuição reflete tanto a natureza das palavras-chave utilizadas na coleta quanto a manifestação linguística de cada sintoma nas postagens.

No presente trabalho, as instâncias do *corpus* Amive são utilizadas como exemplos da classe positiva (textos que expressam sintomas de depressão) nas tarefas de classificação binária. Embora o pipeline proposto possa ser aplicado a cada um dos 18 sintomas anotados no *corpus* de forma independente, os experimentos apresentados neste trabalho concentram-se no sintoma de Tristeza/Humor depressivo, o mais frequente do *corpus*, no qual o classificador é treinado para distinguir postagens que expressam esse sintoma de postagens que não o expressam.

Tabela 2 – Distribuição de instâncias anotadas por sintoma no *corpus* Amive (WILKENS et al., 2026)

Índice	**Sintoma**	**Instâncias**
1	Agitação ou inquietação	15
2	Déficit de atenção ou memória	16
3	Alteração de peso ou apetite	17
4	Alteração de sono	31
5	Sintoma físico	31
6	Dificuldade para decidir	35
7	Sentimento de vazio	40
8	Perda ou diminuição do prazer ou da libido	43
9	Sentimento de culpa	73
10	Irritação ou agressividade	149
11	Alteração na eficiência ou funcionalidade	155
12	Cansaço, desânimo ou fadiga	153
13	Desesperança	164
14	Preocupação, medo ou ansiedade	210
15	Desvalia ou baixa autoestima	234
16	Suicídio ou auto-extermínio	278
17	Desamparo, prejuízo social ou solidão	377
18	Tristeza ou humor depressivo	469
Total		**2.460**

Para compor a classe negativa nas tarefas de classificação, é utilizado um *dataset*

de controle formado por sentenças do *corpus* Amive para as quais nenhum sintoma foi anotado. Para cada experimento, são amostradas aleatoriamente instâncias de controle em quantidade igual ao número de instâncias positivas do sintoma em questão, garantindo um *dataset* balanceado (50% positivo, 50% negativo). A divisão entre treino e teste é feita por identificador de instância, assegurando que textos do mesmo autor não apareçam simultaneamente em ambos os conjuntos.

4.2.2 Ferramentas

Os experimentos foram implementados em Python. Para o *pipeline* de geração de paráfrases, foram utilizadas a biblioteca Transformers (WOLF et al., 2020), que fornece acesso aos modelos PEGASUS e BERTimbau, e a biblioteca `bert-score` (ZHANG et al., 2020b), responsável pelo cálculo da similaridade semântica entre paráfrases candidatas e textos originais. As traduções automáticas foram realizadas por meio do modelo Llama 3.3 70B (AI@Meta, 2024), acessado via API Groq[2]. Para a etapa de classificação, foram utilizadas as implementações de Regressão Logística, SVM, Naive Bayes e Random Forest disponíveis na biblioteca `Scikit-learn` (PEDREGOSA et al., 2011), em conjunto com o vetorizador TF-IDF da mesma biblioteca. A distribuição de categorias gramaticais dos *corpora* original e sintético foi obtida com o spaCy (HONNIBAL et al., 2020), biblioteca de código aberto para Processamento de Linguagem Natural, utilizando o modelo `pt_core_news_lg` treinado para o português brasileiro. O gerenciamento de dados tabulares foi feito com Pandas (The Pandas Development Team, 2020), e o código completo será disponibilizado no GitHub do laboratório LALIC, em <https://github.com/LALIC-UFSCar/synthetic-data-mental-health>.

[2] <https://groq.com>

5 Experimentos e Resultados

Este capítulo apresenta e discute os resultados obtidos para o sintoma de Tristeza/Humor depressivo, o mais frequente no *corpus* Amive, como justificado na subseção 4.2.1.

5.1 Geração dos Dados Sintéticos

5.1.1 *Back-Translation*

A geração dos dados sintéticos foi realizada por meio de uma estratégia de *back-translation* combinada com geração de paráfrases. Assim, o objetivo desse processo é produzir novas versões textuais semanticamente equivalentes aos conteúdos originais, preservando as informações relevantes para tarefas de classificação enquanto introduz variações linguísticas capazes de aumentar a diversidade do conjunto de dados e torná-los não identificáveis.

O *pipeline* de geração recebe como entrada o *corpus* Amive no formato CSV e produz um *dataset* sintético com exatamente uma paráfrase para cada instância do *dataset* original, mantendo os mesmos identificadores e rótulos de sintoma.

A tradução PT→EN e EN→PT foi realizada pelo modelo Llama 3.3 70B via API Groq, com processamento em lote: múltiplos textos são enviados em uma única requisição, numerados sequencialmente, e as traduções extraídas por expressão regular. O tamanho do lote foi definido em 5 textos por requisição, buscando equilibrar eficiência e robustez. Essa configuração reduz o número de chamadas à API, processo que representa potencial gargalo no fluxo, além de diminuir o impacto no caso de falhas. Fora isso, a utilização de lotes menores também reduz o risco de erros de interpretação (parse) na resposta, aumentando a confiabilidade. Entre requisições, um intervalo de 1 segundo é inserido para respeitar os limites de taxa da API.

5.1.2 Geração de Paráfrases

Para cada texto traduzido ao inglês, o modelo PEGASUS foi responsável por gerar 5 paráfrases candidatas. Assim, é utilizada a estratégia *Diverse Beam Search* (VIJAYAKUMAR et al., 2016), configurada com cinco feixes (*beams*) e cinco grupos de feixes (*beam groups*), de forma que cada sequência retornada seja produzida por um caminho de busca distinto. Isto é, como cada grupo contém apenas um feixe, cada candidato é gerado de forma relativamente independente dos demais. Adicionalmente, é aplicada uma penalidade

de diversidade igual a 1.0, reduzindo a pontuação de sequências muito semelhantes às já produzidas por outros grupos e, consequentemente, incentivando o modelo a explorar diferentes formulações linguísticas.

O número de feixes foi fixado em cinco visando balancear qualidade e custo computacional. Em geral, valores maiores ampliam o espaço de busca e aumentam a probabilidade de encontrar sequências mais adequadas, porém exigem mais memória e tempo de processamento. Como o objetivo do trabalho é gerar múltiplos candidatos para posterior seleção automática de apenas um resultado final, cinco feixes mostraram-se suficientes para produzir alternativas diversificadas sem tornar o processo excessivamente custoso.

O comprimento máximo da instância fornecida para ser parafraseadas foi definido em 60 tokens, valor correspondente ao limite de entrada adotado pelo modelo PEGASUS. Como o modelo opera em inglês, a verificação do tamanho dos textos foi realizada sobre as versões já traduzidas para esse idioma, imediatamente antes da etapa de parafraseamento. Essa decisão foi adotada porque a tokenização de textos em português frequentemente produz uma quantidade de tokens diferente daquela obtida para as respectivas traduções em inglês, por não conhecer, por exemplo, caracteres especiais. Dessa forma, a contagem foi realizada sobre a representação efetivamente utilizada como entrada do PEGASUS, garantindo que a restrição de comprimento refletisse com precisão as condições reais de processamento do modelo.

Para os textos cuja versão em inglês ainda excedia esse limite, foi empregada uma estratégia de segmentação hierárquica. Inicialmente, o texto era dividido com base em marcadores de fim de sentença (ponto final, de interrogação ou exclamação) e, quando necessário, fragmentos ainda acima do limite eram subdivididos utilizando vírgulas como separadores adicionais. O procedimento adotado considerava que, caso permanecessem fragmentos acima de 60 tokens após as etapas de segmentação, a instância seria integralmente descartada do processo de geração de paráfrases. Essa escolha evita a introdução de ruídos decorrentes da utilização de textos truncados ou incompletos e foi considerada adequada por se tratar de uma situação esperada como pouco frequente. Na prática, entretanto, nenhuma instância do conjunto positivo precisou ser removida por esse motivo, uma vez que todas puderam ser adequadamente processadas dentro do limite estabelecido.

Após a geração das paráfrases para cada fragmento válido, os resultados eram concatenados automaticamente novamente seguindo a ordem original do texto. Durante essa reconstrução, os sinais de pontuação utilizados na segmentação eram preservados, de forma a manter a estrutura discursiva da postagem. Além disso, eventuais marcas de pontuação finais introduzidas automaticamente pelo PEGASUS eram removidas dos fragmentos intermediários, evitando a inserção de pontuação duplicada durante a junção. O texto reconstruído resultante era então tratado como uma única instância e encaminhado

para a etapa seguinte de tradução de volta ao português.

Esse procedimento foi aplicado apenas aos exemplos da classe positiva. Para a classe de controle, que já possuía quantidade suficiente de instâncias para os experimentos realizados, optou-se por utilizar exclusivamente os textos que originalmente se encontravam dentro do limite de 60 tokens. Dessa forma, evitou-se a inclusão de etapas adicionais de tratamento e possíveis fontes de variabilidade desnecessárias em um conjunto que não demandava aumento de dados sintéticos.

5.1.3 Processamento dos Resultados

Antes de prosseguir com a seleção, as paráfrases candidatas passam por uma etapa de filtragem: são descartadas duplicatas (normalizadas em minúsculas e sem pontuação final) e paráfrases cujo conjunto de tokens é idêntico ao do texto original, garantindo que sejam consideradas apenas reformulações válidas. Tais informações e estatísticas sobre o processo de filtragem são registradas, incluindo a quantidade de duplicatas removidas, a ocorrência de candidatos excessivamente semelhantes ao texto original e medidas de diferença lexical entre o texto de origem e cada paráfrase preservada.

Após a filtragem, o BERTScore F1 é calculado usando o BERTimbau e considerando cada candidata e o texto original em português. A paráfrase com maior pontuação é selecionada. Em caso de empate (duas ou mais candidatas com o mesmo BERTScore F1), a seleção recai sobre a primeira paráfrase na ordem em que foi gerada pelo PEGASUS. Essa ordem corresponde à sequência de feixes de maior probabilidade acumulada na busca em feixe com diversidade. Assim, ainda que as paráfrases empatadas sejam semanticamente equivalentes do ponto de vista do BERTScore, a primeira tende a ser a de maior fluência segundo o modelo, por ter sido gerada pelo feixe mais provável. Caso todas as candidatas sejam descartadas pelos filtros de duplicatas e similaridade com o original, o comportamento padrão adotado é deixar a instância vazia no *dataset* sintético, evitando que o texto original (e portanto os dados sensíveis que ele contém) seja vazado para o *corpus* distribuível. Essa decisão de design prioriza a garantia de privacidade sobre a completude do *corpus* sintético: como a análise prévia do *corpus* indicava que esse cenário seria pouco frequente, optou-se por aceitar a eventual perda de uma pequena fração de instâncias em troca da certeza de que nenhum dado sensível seria exposto inadvertidamente. O número total de instâncias efetivamente excluídas do *corpus* final por esse e outros critérios de filtragem é reportado na subseção 5.1.4.

O *pipeline* implementa também um mecanismo de *checkpoint*: os IDs já processados são registrados em um arquivo JSONL de estatísticas, permitindo retomar o processamento em caso de interrupção sem reprocessar instâncias já concluídas. Esse mecanismo foi especialmente relevante dado o volume do *corpus* e o custo por requisição à API.

5.1.4 Estatísticas do Processo de Geração

Para caracterizar quantitativamente o processo de geração descrito nas seções anteriores, foi implementado um script de auditoria que consolida as estatísticas registradas ao longo do pipeline. Tanto do processamento principal (subseção 5.1.2) quanto do reprocessamento de instâncias cuja tradução ao inglês excedia o limite de 60 tokens, e da substituição de paráfrases selecionadas com BERTScore extremo, descrita a seguir.

Durante a inspeção dos resultados, identificou-se um subconjunto de instâncias cuja paráfrase selecionada apresentava BERTScore F1 igual a 0,0 ou 100,0. Esses valores extremos sinalizam, respectivamente, falha no processo de geração (paráfrase vazia – caso onde o texto era composto apenas por *emojis*, por exemplo – ou completamente dissimilar do original) ou ausência efetiva de reformulação (paráfrase idêntica ao original do ponto de vista semântico), violando a premissa de diversidade textual que orienta este trabalho. Para tratar esses casos, um script de substituição percorreu as demais paráfrases candidatas já submetidas aos filtros de duplicidade e similaridade (subseção 5.1.3) e selecionou, entre as que ainda não haviam sido descartadas, aquela com maior BERTScore estritamente entre 0 e 100. Caso nenhuma candidata satisfizesse essa condição, a instância era removida do *corpus* sintético final sem substituição, pelos mesmos motivos descritos anteriormente: manter um texto vazio ou um substituto sem reformulação genuína comprometeria a garantia de privacidade que orienta as decisões de design do *pipeline*.

A Tabela 3 sintetiza os resultados obtidos para a classe positiva (sintoma de Tristeza/Humor depressivo) e para a classe de controle, correspondentes às mesmas instâncias utilizadas na Tabela 4 e Tabela 5 e na Figura 6, Figura 7 e Figura 8.

Tabela 3 – Estatísticas do processo de geração do corpus sintético

Métrica	**Classe positiva**	**Classe de controle**
Instâncias processadas	446	632
Instâncias mantidas no *corpus* final	445 (99,8%)	630 (99,7%)
Instâncias excluídas	1 (0,2%)	2 (0,3%)
Instâncias substituídas (score 0 ou 100)	43	23
Origem: pipeline principal	401	630
Origem: reprocessamento (textos longos)	44	—
BERTScore F1 (média)	75,90	74,66
BERTScore F1 (mediana)	75,92	75,65
BERTScore F1 (desvio padrão)	8,56	11,66
Diferença lexical, índice de Jaccard (média)	53,71%	55,48%
Diferença lexical, índice de Jaccard (mediana)	56,25%	55,56%
Tokens no texto original (média ± dp)	54,18 ± 43,30	33,50 ± 21,64
Tokens no texto sintético (média ± dp)	39,88 ± 34,22	24,81 ± 13,95

Fonte: Elaborado pela autora.

A taxa de sobrevivência das instâncias foi alta em ambas as classes (99,8% na

positiva e 99,7% no controle), indicando que, na grande maioria dos casos, o pipeline foi capaz de produzir ao menos uma paráfrase candidata com similaridade semântica adequada ao texto original. A necessidade de substituição por score extremo, no entanto, afetou uma parcela não desprezível das instâncias: 43 casos (9,7%) na classe positiva e 23 (3,7%) no controle precisaram ter sua paráfrase final substituída. A maior incidência de substituições na classe positiva é consistente com a natureza desses textos (relatos de sintomas tendem a empregar vocabulário mais específico e menos substituível por sinônimos - conforme discutido na seção 5.5), o que aumenta a chance de o PEGASUS gerar paráfrases excessivamente próximas do original (score 100).

A classe positiva foi a única a passar pela etapa de reprocessamento por segmentação (subseção 5.1.2); 44 das 445 instâncias finais (9,9%) tiveram origem nesse fluxo alternativo, em que o texto é dividido em sentenças e, se necessário, em fragmentos menores antes da paráfrase. No total, 307 fragmentos foram parafraseados com sucesso ao longo desse processo, numa média de 6,98 fragmentos por instância reprocessada, e nenhum fragmento precisou ser descartado por exceder o limite de tokens mesmo após a segmentação. Este resultado confirma a adequação da estratégia de segmentação hierárquica (por pontuação final e, subsidiariamente, por vírgulas).

Em relação à qualidade semântica das paráfrases mantidas, o BERTScore médio foi de 75,90 na classe positiva e 74,66 no controle, com medianas próximas às médias (75,92 e 75,65, respectivamente), o que sugere distribuições relativamente simétricas. O desvio padrão, no entanto, foi consideravelmente maior na classe de controle (11,66 contra 8,56 na positiva), indicando maior variabilidade na qualidade das paráfrases desse grupo, resultado coerente com a maior liberdade de reformulação observada qualitativamente nos textos de controle (seção 5.5), que, por não conterem vocabulário clínico específico, permitem ao modelo de paráfrase um espaço maior de variação tanto para melhor quanto para pior.

A diferença lexical entre original e paráfrase, medida pelo índice de Jaccard sobre os conjuntos de tokens, foi de 53,71% (mediana 56,25%) na classe positiva e 55,48% (mediana 55,56%) no controle. Estes valores são consideravelmente próximos entre si e indicam que, em média, pouco mais da metade do vocabulário de cada par original-paráfrase não é compartilhado entre as duas versões, evidenciando reformulação substancial em ambas as classes.

Por fim, observa-se uma redução expressiva no comprimento médio dos textos entre o original e o sintético: de 54,18 para 39,88 tokens na classe positiva (uma redução de 26,4%) e de 33,50 para 24,81 tokens no controle (redução de 26,0%). Embora o pipeline passe por uma etapa intermediária em inglês, a comparação de comprimento entre o texto original em português e sua tradução não é diretamente interpretável: os dois idiomas possuem estruturas sintáticas distintas que naturalmente produzem sentenças de comprimentos diferentes, de modo que um texto mais curto em inglês pode perfeitamente resultar em

uma retradução de comprimento equivalente ou superior ao original em português. Dada essa dificuldade de isolar a contribuição da tradução, a redução observada é atribuída principalmente à etapa de parafraseamento com o PEGASUS, cujo papel é discutido em detalhe na subseção 5.5.1: ao gerar paráfrases de sentenças com múltiplas orações coordenadas e subordinadas (estrutura típica do gênero desabafo), o modelo tende a podar conteúdo em vez de preservar integralmente a estrutura do texto de origem.

Cabe notar que as sentenças da classe de controle já eram, no *corpus* original, naturalmente mais curtas do que as da classe positiva (33,50 contra 54,18 tokens em média), o que decorre diretamente do critério de seleção descrito na subseção 5.1.2: como a classe de controle já possuía instâncias suficientes para os experimentos realizados, optou-se por utilizá-la apenas com os textos que originalmente se encontravam dentro do limite de 60 tokens do PEGASUS, evitando a etapa adicional de segmentação. A classe positiva, por sua vez, inclui também as instâncias mais longas, processadas pelo fluxo de reprocessamento descrito anteriormente. Ainda assim, a redução proporcional observada entre original e sintético foi praticamente idêntica nas duas classes (26,4% e 26,0%), o que sugere uma causa comum subjacente ao pipeline de paráfrase, e não uma particularidade do conteúdo clínico da classe positiva ou do critério de seleção que diferencia as duas classes.

Essa redução não decorre do limite máximo de tokens aceito pelo PEGASUS como entrada (uma vez que os textos mais longos já são segmentados em unidades menores antes da etapa de paráfrase), mas sim da dificuldade do modelo em manter coerência referencial e completude informacional ao processar textos com múltiplas orações coordenadas e subordinadas, estrutura comum ao gênero *desabafo* característico de postagens em redes sociais sobre saúde mental. Em outras palavras, ao parafrasear sentenças sintaticamente complexas, o PEGASUS tende a podar conteúdo (omitindo orações, fundindo ideias ou eliminando informação secundária) em vez de preservar integralmente a estrutura proposicional do texto de origem.

Como direção de mitigação para trabalhos futuros, sugere-se a segmentação de textos longos em unidades sentenciais menores antes da etapa de paráfrase, de modo a reduzir a probabilidade de poda de conteúdo. Vale notar que, ao realizar tal processo, deve-se avaliar qual seu impacto na perda semântica da separação de orações com conteúdo relacionado/dependente.

5.2 Treinamento dos Classificadores

5.2.1 Construção e Organização dos *Datasets*

Para cada sintoma, são construídos dois *datasets* binários independentes: um com as instâncias originais do *corpus* Amive e outro com as instâncias sintéticas geradas na

etapa anterior. Em ambos os casos, as instâncias do sintoma de interesse recebem rótulo positivo (1, presente), e instâncias de controle (postagens sem indicação de sintomas) recebem rótulo negativo (0, ausente).

As instâncias de controle são amostradas aleatoriamente em quantidade igual ao número de instâncias positivas do sintoma em questão, garantindo um *dataset* balanceado (50% positivo, 50% negativo). Tanto para os exemplos positivos quanto para os de controle, são utilizadas versões originais e sintéticas associadas aos mesmos identificadores, permitindo a construção de conjuntos equivalentes nos diferentes cenários experimentais.

A divisão entre conjuntos de treino e teste é feita por identificador de instância, seguindo a proporção 70%/30% – convenção amplamente adotada em experimentos de classificação de texto (MANNING; RAGHAVAN; SCHÜTZE, 2008), que reserva uma fração suficiente de dados para avaliação sem comprometer o tamanho do conjunto de treinamento. A divisão é feita no *dataset* original e replicada no sintético: os mesmos IDs de treino e teste são usados nos dois cenários, garantindo que a comparação seja feita sobre exatamente os mesmos exemplos de teste.

A semente aleatória foi fixada em 42 em todas as operações que envolvem aleatoriedade (amostragem do controle, divisão treino/teste e inicialização dos classificadores) garantindo reprodutibilidade integral dos experimentos.

5.2.2 Vetorização e Classificação

As representações TF-IDF resultantes são utilizadas como entrada para os três classificadores descritos na subseção 4.1.2.2: Regressão Logística, SVM linear, Naive Bayes e Random Forest. As configurações dos modelos foram definidas com base em boas práticas consolidadas para tarefas de classificação de texto representadas por TF-IDF, buscando garantir estabilidade e comparabilidade experimental entre os classificadores.

Para a Regressão Logística, o número máximo de iterações foi fixado em 1000, valor adotado para favorecer a convergência do algoritmo de otimização numérica. Em problemas de alta dimensionalidade, como aqueles gerados por representações TF-IDF, é comum que valores padrão inferiores não sejam suficientes para atingir convergência, especialmente quando o espaço de atributos é esparso e contém grande número de termos.

Para o SVM com kernel linear, a escolha do kernel linear é justificada pela própria natureza da representação TF-IDF e seu espaço vetorial de alta dimensionalidade. Nesse contexto, separadores lineares tendem a apresentar desempenho competitivo em relação a kernels não lineares, com a vantagem adicional de menor custo computacional e maior escalabilidade.

Para o Naive Bayes, foi utilizada a variante Multinomial, amplamente empregada em tarefas de classificação de texto com representações TF-IDF. Esse modelo assume

independência condicional entre os atributos e é particularmente eficiente em cenários de alta dimensionalidade e dados esparsos, como os vetores de frequência ponderada por termos. Apesar de sua simplicidade, o Naive Bayes frequentemente apresenta desempenho competitivo em problemas de classificação textual, servindo como um forte *baseline* probabilístico para comparação com modelos discriminativos.

Já o Random Forest foi incluído como um classificador baseado em conjunto de árvores de decisão, configurado com parâmetros padrão da biblioteca utilizada. Esse modelo introduz uma abordagem não linear ao problema, permitindo capturar interações entre termos que não são modeladas pelos classificadores lineares. Além disso, o uso de múltiplas árvores reduz a variância do modelo e aumenta sua robustez em relação a variações no conjunto de treinamento, o que o torna adequado para avaliar a consistência dos efeitos da inclusão de dados sintéticos em diferentes regimes de aprendizado.

A utilização de múltiplos classificadores tem como objetivo reduzir a dependência dos resultados em relação a um único algoritmo de aprendizado. Dessa forma, torna-se possível avaliar se os efeitos da inclusão dos dados sintéticos se mantêm consistentes em modelos com diferentes características de aprendizado. Todos os classificadores foram treinados e avaliados sobre a mesma representação vetorial e sob as mesmas divisões de treino e teste, garantindo comparabilidade entre os resultados obtidos.

5.3 Avaliação dos Classificadores

O objetivo deste experimento é verificar em que medida o *corpus* sintético preserva a utilidade do *corpus* original para a tarefa de classificação binária, avaliando se classificadores treinados com dados sintéticos atingem desempenho equivalente nas métricas apresentadas na subseção 2.4.1 ao de classificadores treinados com dados reais.

Os resultados apresentados nesta seção referem-se ao sintoma de Tristeza/Humor depressivo (sintoma 18), o mais frequente no *corpus* Amive com 469 instâncias anotadas. Os valores obtidos serão primeiramente descritos para, em seguida, serem analisados e compreendidos. A Tabela 4 e a Tabela 5 apresentam os valores de acurácia, precisão, revocação e F1 obtidos pelos quatro classificadores nos quatro cenários: original (Orig.), sintético (Sint.) e cruzados (S→O e O→S). O cenário S→O é particularmente relevante para o objetivo desta pesquisa, pois simula a situação de uso real do *corpus* sintético: um classificador treinado exclusivamente com dados sintéticos sendo avaliado sobre dados reais, sem acesso aos textos originais.

5.3.1 Análise por Métrica

A Figura 6 apresenta os resultados agrupados por métrica, permitindo comparar o comportamento dos quatro classificadores em cada dimensão de avaliação e em cada um

Tabela 4 – Resultados dos classificadores para o sintoma Tristeza/Humor depressivo – Acurácia e Precisão

Modelo	Acurácia (Orig.)	Acurácia (Sint.)	Acurácia (S→O)	Acurácia (O→S)	Precisão (Orig.)	Precisão (Sint.)	Precisão (S→O)	Precisão (O→S)
Logistic Regression	0.8127	0.8315	0.779	0.8127	0.8358	0.8682	0.7872	0.8689
SVM	0.839	0.8352	0.7828	0.8165	0.854	0.8582	0.7628	0.8527
Naive Bayes	0.7191	0.779	0.7116	0.7228	0.6684	0.7314	0.6649	0.6701
Random Forest	0.7978	0.824	0.8202	0.8015	0.8258	0.8974	0.8651	0.8718

Tabela 5 – Resultados dos classificadores para o sintoma Tristeza/Humor depressivo – Recall e F1

Modelo	Recall (Orig.)	Recall (Sint.)	Recall (S→O)	Recall (O→S)	F1 (Orig.)	F1 (Sint.)	F1 (S→O)	F1 (O→S)
Logistic Regression	0.8	0.8	0.7929	0.7571	0.8175	0.8327	0.79	0.8092
SVM	0.8357	0.8214	0.85	0.7857	0.8448	0.8394	0.8041	0.8178
Naive Bayes	0.9214	0.9143	0.9071	0.9286	0.7748	0.8127	0.7674	0.7784
Random Forest	0.7786	0.75	0.7786	0.7286	0.8015	0.8171	0.8195	0.7938

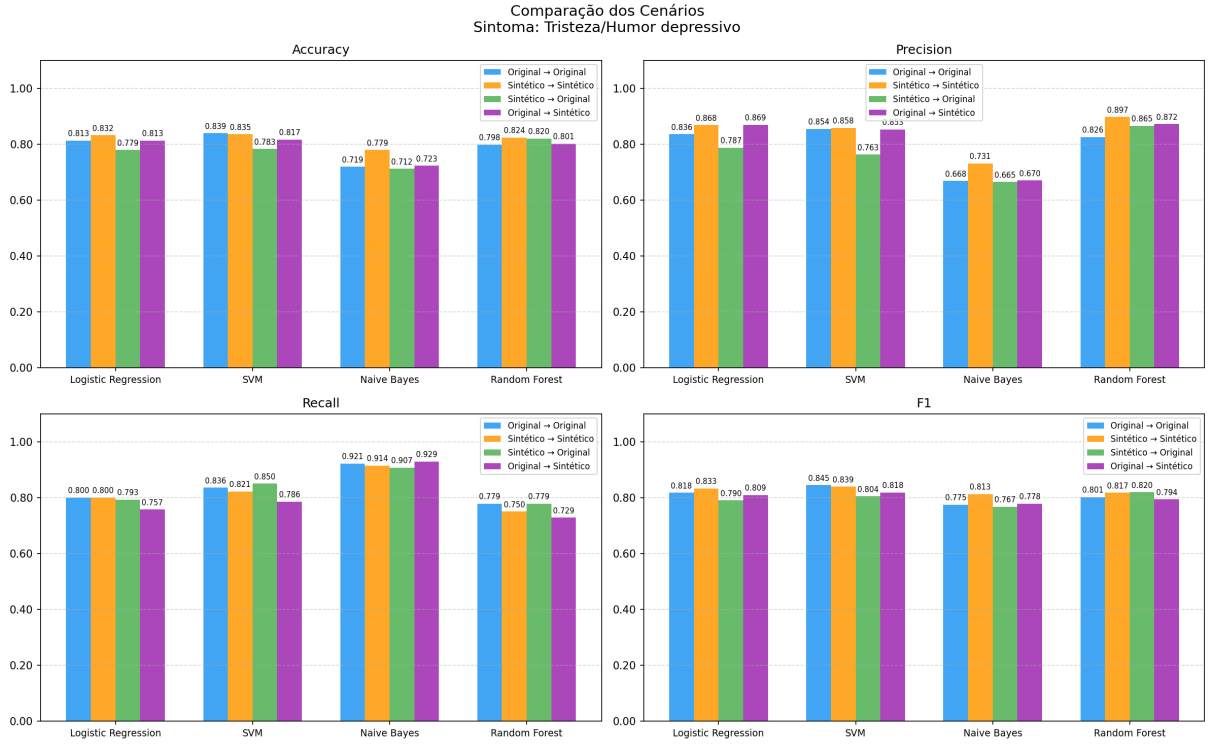
dos quatro cenários. Em acurácia, o cenário Sintético → Sintético supera ou se iguala o Original → Original em todos os modelos, com ganhos mais expressivos para Naive Bayes (0,719 para 0,779) e Random Forest (0,798 para 0,824). Isso indica que o *corpus* sintético constitui um conjunto de dados internamente coerente, sobre o qual os classificadores conseguem aprender padrões discriminativos ao menos tão eficazmente quanto sobre os dados originais. O cenário Sintético → Original, por outro lado, apresenta queda em relação ao *baseline* original para todos os modelos, sendo a Regressão Logística (0,813 para 0,779) e o SVM (0,839 para 0,783) os mais afetados – o que indica que, embora o *corpus* sintético seja internamente coerente, os classificadores treinados exclusivamente com paráfrases ainda enfrentam dificuldades para generalizar para o vocabulário e as estruturas linguísticas dos textos reais.

Em precisão, o mesmo padrão se repete de forma mais acentuada: o cenário Sintético → Sintético melhora consistentemente em relação ao original, com destaque para Random Forest (0,826 para 0,897) e Naive Bayes (0,668 para 0,731). Já o cenário Sintético → Original apresenta as maiores quedas observadas neste experimento, sendo o SVM o mais impactado (0,854 para 0,763, uma redução de quase 0,09). Uma possível explicação para essa maior sensibilidade do SVM está na natureza de sua fronteira de decisão, definida exclusivamente pelos vetores de suporte mais próximos da margem, e não pela distribuição completa dos dados de treino, como ocorre na Regressão Logística. Assim, a variação lexical introduzida pelo processo de paráfrase, ainda que moderada, pode deslocar de forma mais acentuada quais instâncias passam a atuar como vetores de suporte, alterando significativamente a margem aprendida quando o modelo é avaliado sobre o vocabulário real do conjunto de teste original. Essa hipótese se mostra consistente com a robustez relativamente maior observada no Random Forest nesse mesmo cenário, cuja agregação de múltiplas árvores sobre subamostras distintas de atributos tende a atenuar o impacto de variações pontuais no vocabulário.

Em revocação, o comportamento diverge dos demais indicadores. O cenário Sintético

→ Original mantém-se estável ou até melhora em relação ao original para a maioria dos modelos: SVM (0,773 para 0,850) e Naive Bayes (0,921 para 0,907, queda mínima), enquanto o cenário Original → Sintético tende a apresentar quedas mais visíveis, como na Regressão Logística (0,800 para 0,757) e no Random Forest (0,779 para 0,729). O F1, que sintetiza precisão e revocação, reflete o equilíbrio entre essas tendências: o cenário Sintético → Sintético é o mais consistentemente positivo em relação ao original, enquanto o cenário Sintético → Original apresenta o padrão mais heterogêneo, com perdas em modelos lineares e estabilidade relativa no Random Forest.

Figura 6 – Comparação dos resultados por métrica entre os diferentes cenários avaliados



5.3.2 Análise por Modelo

A Figura 7 reorganiza os mesmos dados agrupando por modelo, facilitando a visualização do perfil de desempenho de cada classificador ao longo dos quatro cenários avaliados. A Regressão Logística apresenta o cenário Sintético → Sintético como o mais favorável em todas as métricas, com destaque para a precisão (0,836 para 0,868). No entanto, o cenário Sintético → Original revela uma queda generalizada em relação ao *baseline* original, com a revocação sendo a métrica mais resiliente (0,800 para 0,793) – sugerindo que, mesmo com dificuldades de generalização para o vocabulário real, o modelo mantém razoável capacidade de identificar corretamente as instâncias positivas.

O SVM exibe um padrão semelhante ao da Regressão Logística para os cenários Sintético → Sintético, mas se diferencia no cenário Sintético → Original: enquanto a

precisão sofre a maior queda observada no experimento (0,854 para 0,763), a revocação na verdade melhora (0,773 para 0,850), atingindo o maior valor absoluto entre os quatro modelos nesse cenário. Esse comportamento indica que, mesmo perdendo capacidade de evitar falsos positivos, o SVM treinado com dados sintéticos se torna mais sensível para identificar corretamente as instâncias positivas reais.

O Naive Bayes apresenta o perfil mais estável entre os quatro modelos: as diferenças entre os cenários são pequenas em todas as métricas, e a revocação se mantém consistentemente alta (acima de 0,90) em todos os cenários, incluindo Sintético → Original (0,907). Esse comportamento é coerente com a natureza probabilística do Naive Bayes, que tende a ser mais robusto a variações na distribuição de frequência de termos do que classificadores baseados em fronteiras de decisão rígidas.

O Random Forest, por sua vez, mantém acurácia e F1 relativamente estáveis entre o cenário original e o cenário Sintético → Original (0,798 para 0,820 em acurácia), com a precisão inclusive melhorando (0,826 para 0,865). A revocação, no entanto, é a métrica mais penalizada nesse cenário (0,779 para 0,778, estável) e principalmente no cenário Original → Sintético (0,779 para 0,729), sugerindo que o modelo treinado em dados reais tem mais dificuldade para generalizar corretamente para o conjunto sintético do que o inverso.

De modo geral, observa-se que o cenário Sintético → Sintético tende a favorecer todos os classificadores de forma consistente, enquanto o cenário Sintético → Original (o mais relevante para validar a utilidade prática do *corpus* sintético) apresenta um comportamento mais heterogêneo entre os modelos: queda acentuada em precisão para os modelos lineares, mas preservação ou ganho de revocação, especialmente no SVM e no Naive Bayes.

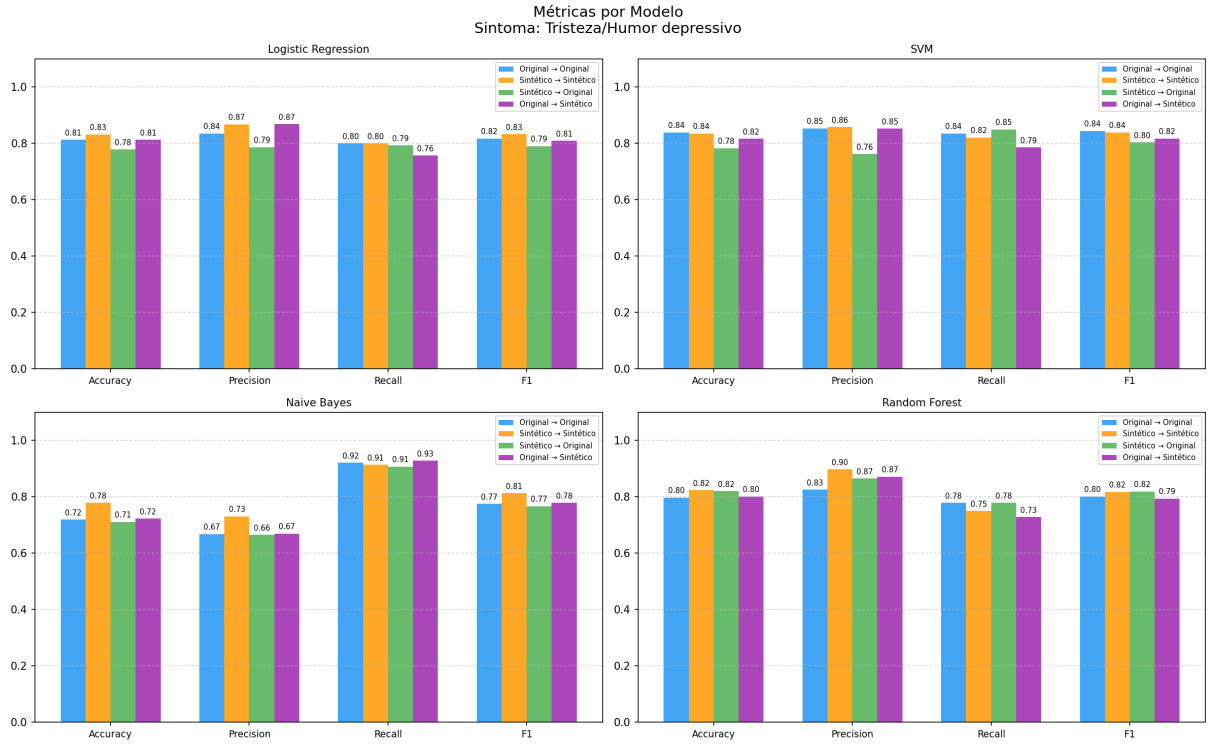
5.3.3 Mapa de Calor

O mapa de calor da Figura 8 sintetiza diretamente as diferenças entre os cenários Sintético → Sintético, Sintético → Original e Original → Sintético em relação ao *baseline* Original → Original, para cada combinação de modelo e métrica. Células verdes indicam melhoria, células laranjas indicam queda, e a intensidade da cor reflete a magnitude da diferença.

O painel à esquerda (Sintético − Original) é predominantemente verde, confirmando que treinar e testar exclusivamente no *dataset* sintético tende a produzir desempenho igual ou superior ao *baseline* original em todos os quatro modelos, com destaque para os ganhos de precisão do Naive Bayes (+0,063) e do Random Forest (+0,072).

O painel central (Sintético → Original − Original) é o mais relevante para avaliar a utilidade prática do *corpus* sintético, pois representa o cenário em que um classifica-

Figura 7 – Comparação dos resultados por modelo entre os diferentes cenários avaliados



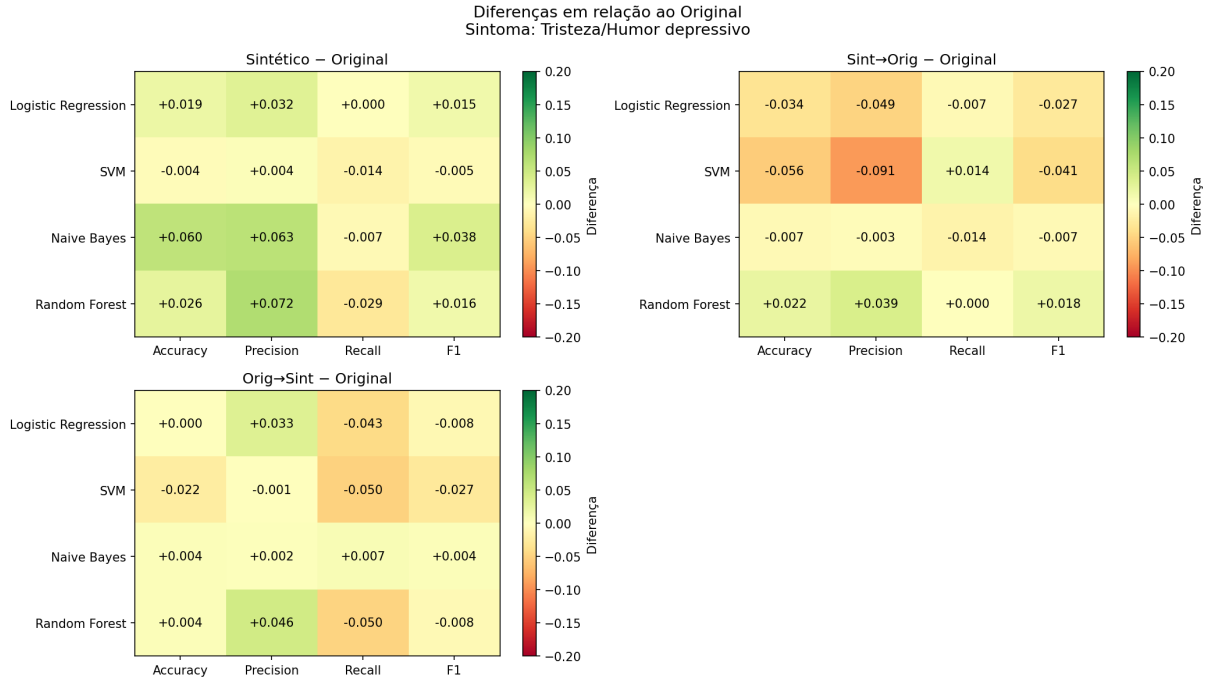
dor treinado apenas com dados sintéticos é aplicado sobre dados reais. Esse painel é predominantemente laranja na coluna de precisão, com as maiores quedas absolutas do experimento: SVM (-0,091) e Regressão Logística (-0,049). Em contraste, a coluna de revocação apresenta o padrão oposto, com ganhos para SVM (+0,014) e Random Forest (+0,000, estável), e queda apenas marginal para Naive Bayes (-0,014). Esse resultado reforça a observação de que o *dataset* sintético tende a deslocar o classificador em direção a uma maior sensibilidade para a classe positiva, custando redução na precisão.

O painel à direita (Original → Sintético − Original) mostra um padrão intermediário: a precisão se mantém relativamente estável ou levemente positiva para a maioria dos modelos, mas a revocação é consistentemente penalizada, com quedas de até -0,050 tanto para SVM quanto para Random Forest. Isso indica que classificadores treinados nos dados originais têm maior dificuldade em recuperar corretamente as instâncias positivas quando avaliados sobre o *dataset* sintético, possivelmente em decorrência da variação lexical introduzida pelas paráfrases.

5.3.4 Síntese

Em conjunto, os resultados da avaliação extrínseca indicam que o *corpus* sintético constitui um substituto funcional do *corpus* original para a tarefa de classificação binária do sintoma de Tristeza/Humor depressivo, ainda que não isento de limitações. O objetivo central desta avaliação (verificar se classificadores treinados com dados sintéticos atingem

Figura 8 – Mapas de calor das diferenças de desempenho em relação ao conjunto original



desempenho equivalente ao de classificadores treinados com dados reais) é parcialmente atingido: no cenário mais relevante (Sintético → Original), a revocação, métrica prioritária em aplicações de triagem em saúde mental, é preservada ou até melhorada na maioria dos modelos, enquanto a precisão apresenta quedas mais consistentes, especialmente nos modelos lineares. Esse padrão sugere que o *corpus* sintético desloca os classificadores em direção a uma maior sensibilidade para a classe positiva, à custa de um aumento nos falsos positivos (*trade-off* que, no contexto de triagem em saúde mental, tende a ser preferível ao inverso).

5.4 Avaliação Intrínseca

5.4.1 Vocabulário, D-2 e BLEU

Em relação à avaliação intrínseca, pelo valor do tamanho do vocabulário apresentado na Tabela 6, nota-se que houve uma diminuição na variação lexical do *corpus* sintético (2067) em relação ao *corpus* original (2116) de cerca de 49 *types* (tokens únicos) para a classe positiva. Esse resultado é consistente com a redução de aproximadamente 26% no comprimento médio dos textos sintéticos observada na subseção 5.1.4: como o PEGASUS tende a podar conteúdo ao parafrasear sentenças longas, os textos sintéticos resultantes são mais curtos e, consequentemente, empregam um número menor de palavras distintas no total. Quanto ao D-2, nota-se uma melhora no valor no *corpus* sintético (0,7243) em relação ao *corpus* original (0,6575), indicando uma diversidade maior de bi-gramas apesar

da leve redução no vocabulário total. Por fim, o baixo valor de BLEU invertido (0,1394 em uma escala de 0 a 1) também é um bom indicativo de qualidade intrínseca: situado próximo ao extremo inferior da escala, indica que as sentenças do *corpus* sintético compartilham poucos n-gramas com as sentenças do *corpus* original, evidenciando reformulação genuína por parte do pipeline de paráfrase.

Tabela 6 – Métricas de qualidade do *corpus* sintético comparado ao original **(classe positiva)**.

Métrica	Original	Sintético
Vocabulário	2116	2067.0
D-2 (bigramas únicos)	0.6575	0.7243
BLEU (invertido)	—	0.1394

Para a classe de controle, apresentada na Tabela 7, observa-se um padrão semelhante, porém mais acentuado. O tamanho do vocabulário também diminui no *corpus* sintético (2246) em relação ao original (2409), uma redução de 163 *types*, proporcionalmente maior do que a observada na classe positiva. O D-2 segue a mesma tendência de melhora (0,7489 para 0,7796), e o BLEU invertido (0,1218) é ainda mais baixo do que o obtido para a classe positiva, sugerindo que as paráfrases geradas para os textos de controle se afastam ainda mais estruturalmente dos textos originais. Esse comportamento é consistente com a discussão da seção 5.4: por não conterem conteúdo emocionalmente marcado ou sintomático, os textos de controle parecem oferecer ao modelo de paráfrase maior liberdade de reformulação.

Tabela 7 – Métricas de qualidade do *corpus* sintético comparado ao original **(classe de controle)**.

Métrica	Original	Sintético
Vocabulário	2409	2246.0
D-2 (bigramas únicos)	0.7489	0.7796
BLEU (invertido)	—	0.1218

5.4.2 Distribuição de categorias gramaticais

A Figura 9 e a Figura 10 apresentam a distribuição de categorias gramaticais (*part-of-speech*, POS) dos *corpora* original e sintético, para as classes positiva e de controle, respectivamente, obtidas com o modelo `pt_core_news_lg` do spaCy.

Para a classe positiva (Figura 9), a distribuição entre original e sintético é relativamente estável: as categorias mais frequentes, substantivos (`NOUN`, de 0,194 para 0,184) e verbos (`VERB`, de 0,178 para 0,191), apresentam pequenas variações em sentidos opostos, mantendo a soma das duas categorias praticamente constante. Os pronomes (`PRON`) apresentam o maior aumento relativo (0,107 para 0,120), enquanto os substantivos próprios

(PROPN) - nomes específicos de pessoas, lugares ou instituições - constituem a categoria mais reduzida em termos absolutos (0,021 para 0,008), uma queda de 62%. Essa redução de substantivos próprios é coerente com o esperado em um pipeline de tradução e paráfrase, que tende a generalizar ou omitir nomes específicos.

Para a classe de controle (Figura 10), a distribuição apresenta divergências mais acentuadas entre original e sintético. O aumento mais expressivo ocorre nos pronomes (PRON, de 0,115 para 0,140, um acréscimo de 0,025), e os advérbios (ADV) apresentam a maior queda entre todas as categorias em ambos os *corpora* (0,089 para 0,072, uma redução de 19%). Assim como na classe positiva, os substantivos próprios (PROPN) sofrem a maior redução relativa (0,036 para 0,013, uma queda de 64%), reforçando a tendência de menor presença de nomes próprios nos textos sintéticos, um resultado favorável do ponto de vista da privacidade, na medida em que reduz a presença de potenciais identificadores residuais no *corpus* distribuído.

De modo geral, a comparação entre a Figura 9 e a Figura 10 indica que a classe de controle sofre uma distorção sintática proporcionalmente maior do que a classe positiva ao longo do pipeline de geração. É importante destacar que o pipeline não possui qualquer mecanismo que reconheça ou priorize conteúdo clínico: tanto a geração de paráfrases quanto a tradução automática tratam cada frase de forma genérica, sem distinção entre textos que descrevem sintomas e textos de conteúdo neutro. Uma hipótese para explicar essa diferença é de natureza puramente estatística: textos da classe positiva tendem a conter um vocabulário mais específico e menos substituível por sinônimos sem alteração perceptível de significado (como termos diretamente associados a estados emocionais ou sintomas), o que levaria o critério de seleção por similaridade semântica global a favorecer, por consequência, paráfrases estruturalmente mais próximas do texto original nesses casos. Já os textos de controle, por não possuírem esse vocabulário mais restrito, permitiriam um espaço maior de reformulações que ainda preservam alta similaridade semântica segundo o BERTScore, resultando em maior variação sintática observada. Essa explicação permanece como hipótese a ser investigada em trabalhos futuros, e não como conclusão estabelecida a partir dos dados aqui apresentados.

5.5 Discussão dos Resultados

5.5.1 Análise Qualitativa

5.5.1.1 Exemplos Representativos

Com o objetivo de avaliar criticamente a qualidade do *corpus* sintético produzido pelo pipeline, foi conduzida uma análise qualitativa de exemplos representativos, selecionados a partir da pontuação BERTScore e de uma inspeção manual do conteúdo clínico

Figura 9 – Distribuição de classes gramaticais do *corpus* sintético comparado ao original **(classe positiva)**

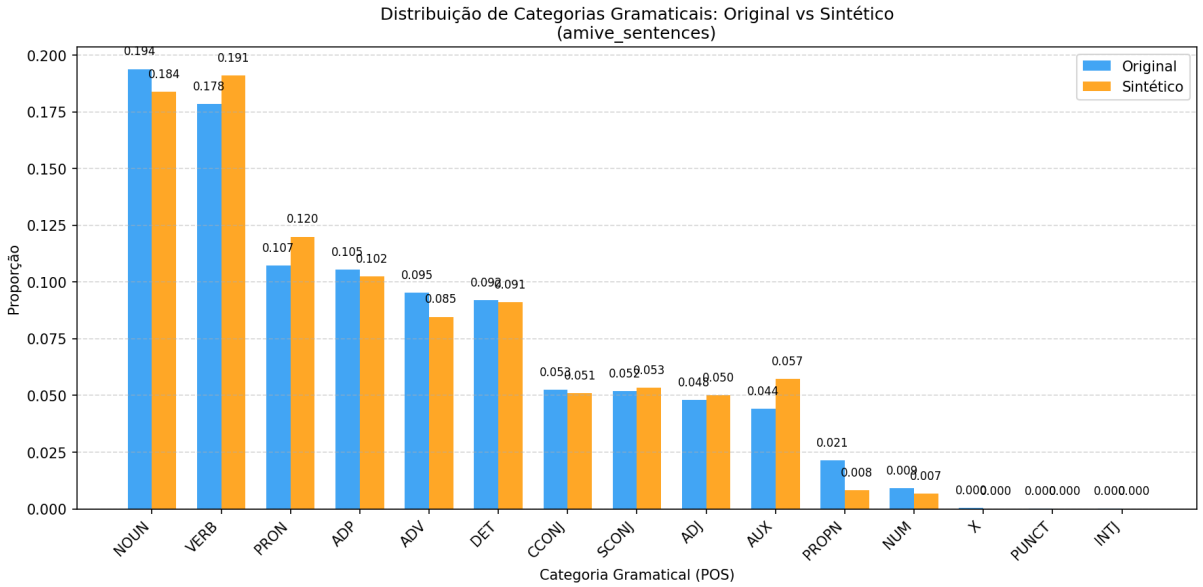
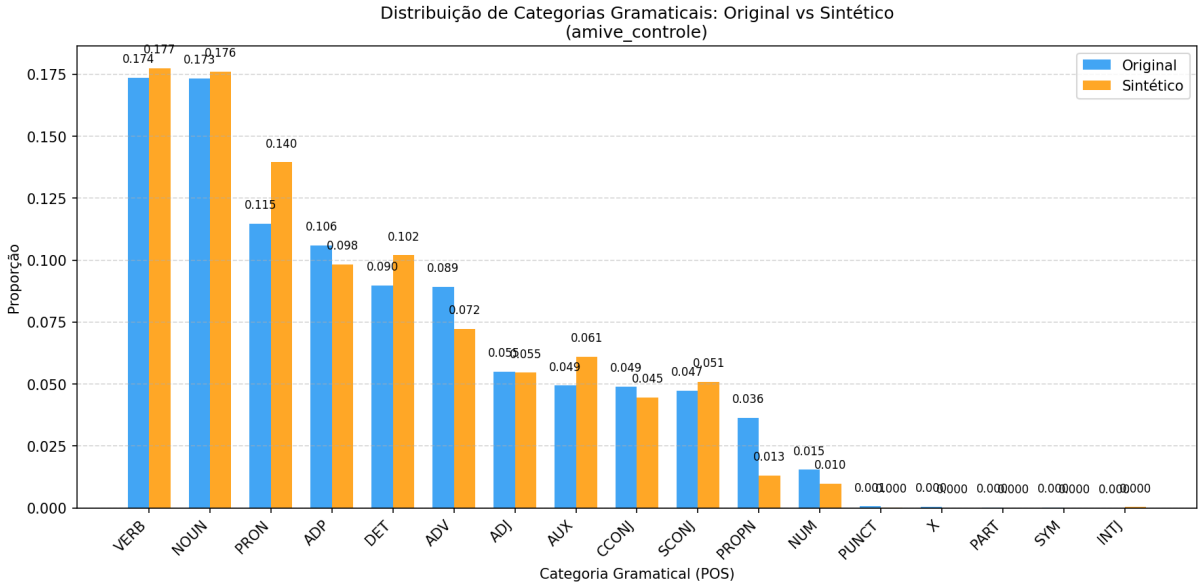


Figura 10 – Distribuição de classes gramaticais do *corpus* sintético comparado ao original **(classe de controle)**



preservado. Para cada exemplo analisado, são apresentados três elementos: (i) a melhor paráfrase obtida em inglês, produto direto da etapa de paráfrase; (ii) a correspondente melhor paráfrase em português (resultado final que será adicionado ao *corpus* sintético), obtida pela retradução do texto parafraseado; e (iii) a pontuação BERTScore associada a essa paráfrase, utilizada como métrica automática de similaridade semântica entre o texto original (em português) e o texto gerado.

Deliberadamente, o texto original em português, sua tradução para o inglês e

o identificador de cada instância são omitidos desta seção, uma vez que o dado bruto corresponde a relatos pessoais sensíveis cuja publicação foi vedada pela própria motivação deste trabalho. Dessa forma, a análise qualitativa aqui apresentada baseia-se exclusivamente em material já transformado pelo pipeline (traduções e paráfrases), preservando o compromisso de não exposição do conteúdo original mesmo no contexto da discussão metodológica do próprio trabalho.

5.5.1.1.1 Preservação Semântica Adequada

O exemplo do Quadro 1 ilustra um caso em que a paráfrase produzida apresenta reescrita com baixa sobreposição lexical em relação ao texto original, mantendo, ao mesmo tempo, a totalidade da informação clinicamente relevante (sintomas e consequência funcional).

Quadro 1 – Exemplo de preservação semântica adequada (*score* = 73,0)

Campo	**Texto**
Melhor paráfrase (EN)	*I dropped out of college because of my anxiety and depression.*
Melhor paráfrase (PT)	Eu abandonei a faculdade por causa da minha ansiedade e depressão.

Nesse caso, o modelo inverteu a estrutura causal da sentença original (causa → efeito) para uma estrutura efeito → causa, sem comprometer a integridade da informação. Os dois sintomas mencionados no texto-fonte (ansiedade e depressão) e a consequência funcional relatada (evasão escolar) permanecem intactos em ambas as línguas, apesar de perder o marcador de intensidade que existia no texto original. A baixa taxa de sobreposição lexical entre o texto original e a paráfrase (aproximadamente 29% das palavras) é um indicador de que a transformação não se limita a substituições pontuais, mas envolveu reorganização sintática, desejável para a geração de dados sintéticos com diversidade textual.

5.5.1.1.2 Preservação Parcial

Em outros casos, a paráfrase preserva o conteúdo semântico central, mas mantém grande parte da estrutura sintática do texto original, sobretudo em trechos entre aspas (citações diretas), o que reduz a diversidade lexical introduzida pelo pipeline.

O exemplo do Quadro 2 preserva tanto o sintoma relatado (fase depressiva) quanto um padrão de discurso clinicamente relevante: a sensação de isolamento social mesmo diante de promessas de apoio por parte de terceiros (um marcador frequentemente associado a relatos de solidão e desesperança em estudos sobre depressão). Contudo, a geração de paráfrases parece ter problema para reformular a citação direta entre aspas, o que sugere que o modelo de paráfrase trata o discurso reportado como um bloco de menor

Quadro 2 – Exemplo de preservação parcial (*score* = 71,02)

Campo	**Texto**
Melhor paráfrase (EN)	*I'm in a bad phase of depression and I'm constantly told "when you need it, just call and I'll be here", but when I needed it, no one was available for me.*
Melhor paráfrase (PT)	Estou em uma fase ruim de depressão e constantemente me dizem "quando precisar, basta ligar e estarei aqui", mas quando precisei, ninguém estava disponível para mim.

maleabilidade sintática. Esse comportamento pode decorrer de um viés aprendido pelo PEGASUS durante seu pré-treinamento, no qual citações diretas tendem a ser reproduzidas literalmente em vez de parafraseadas, refletindo um padrão comum em textos jornalísticos e narrativos. Adicionalmente, o próprio critério de seleção por BERTScore pode reforçar esse efeito: candidatas que preservam a citação inalterada tendem a obter maior similaridade semântica em relação ao texto original, sendo favorecidas na etapa de seleção mesmo quando o restante da sentença foi efetivamente reformulado.

5.5.1.1.3 Perda Semântica Crítica

O caso a seguir (Quadro Quadro 3) representa o cenário mais problemático identificado na análise: a perda completa do token mais relevante para a tarefa de classificação (*depressão*), resultante da literalização de uma construção metafórica.

Quadro 3 – Exemplo de perda semântica crítica (*score* = 41,5)

Campo	**Texto**
Melhor paráfrase (EN)	*This huge backpack is causing everyone to fall down.*
Melhor paráfrase (PT)	Essa mochila enorme está fazendo todo mundo cair.

No texto original, a depressão é personificada por meio de uma metáfora ("mochila gigante" que arrasta as pessoas para baixo). O modelo de paráfrase eliminou o sujeito metafórico (*depressão*) e manteve apenas o predicado literal da metáfora (a mochila fazendo as pessoas caírem), produzindo uma sentença que, isoladamente, não carrega qualquer marcador linguístico associável a sintomas depressivos. Esse padrão de erro é particularmente grave em um *corpus* destinado à classificação de sintomas, pois remove justamente o rótulo semântico-lexical mais discriminativo da amostra.

5.5.1.1.4 Perda de Perspectiva Pessoal em Relatos Autobiográficos

Outro padrão problemático encontrado foi observado em textos que relatam um evento autobiográfico específico (por exemplo, um diagnóstico clínico). Nesses casos, houve instâncias nas quais a paráfrase tendeu a converter o relato em primeira pessoa em uma

proposição genérica e impessoal, alterando o gênero discursivo da amostra e tornando-a majoritariamente descritiva. Um exemplo pode ser observado no Quadro 4.

Quadro 4 – Exemplo de despersonalização do relato (*score* = 72,71)

Campo	**Texto**
Melhor paráfrase (EN)	*Moderate depression is when depression and suicidal thoughts interfere with life outside the home.*
Melhor paráfrase (PT)	A depressão moderada é quando a depressão e os pensamentos suicidas interferem com a vida fora de casa.

Na postagem original havia um autorrelato de diagnóstico de depressão que foi substituído por uma sentença em terceira pessoa com estrutura de definição ("A depressão moderada é quando..."), aproximando-se mais de um enunciado descritivo e explicativo do que de um relato pessoal de sofrimento. Embora o conteúdo proposicional (depressão moderada associada a pensamentos suicidas e interferência na vida fora de casa) permaneça preservado, a alteração da perspectiva do falante é relevante para a tarefa de classificação de sintomas: *corpora* de relatos pessoais costumam carregar marcadores linguísticos de subjetividade (pronomes em primeira pessoa, tempos verbais perfectivos, hesitação) que constituem parte do sinal utilizado por classificadores treinados nesse domínio. A conversão sistemática desses relatos em proposições impessoais pode, portanto, deslocar a distribuição estatística dos dados sintéticos em relação à distribuição dos dados originais.

5.5.1.2 Demais Observações

Além dos padrões anteriormente discutidos, a inspeção do *corpus* revelou outras tendências qualitativas relevantes para a tarefa de classificação de sintomas depressivos. Observou-se, em primeiro lugar, uma tendência de neutralização do registro coloquial do texto original: marcadores de oralidade e informalidade típicos de postagens em redes sociais (como gírias, intensificadores e interjeições) tendem a ser suavizados ou eliminados nas paráfrases geradas, resultando em um texto sintática e lexicalmente mais padronizado do que o original. Esse efeito é relevante porque a atenuação sistemática destes marcadores pode reduzir a força do sinal discriminativo presente nos dados sintéticos.

Observou-se também outro padrão referente à *perda de marcadores de modalidade epistêmica*: expressões que indicam grau de certeza ou hesitação do autor em relação ao próprio relato (como "talvez", "acho que", "não sei se") foram, em diversos casos, omitidas ou convertidas em afirmações categóricas pela paráfrase. Essa alteração é relevante porque a incerteza autoatribuída (por exemplo, "talvez eu seja depressivo") constitui um padrão linguístico distinto de uma afirmação direta ("eu sou depressivo"), podendo representar diferentes estágios de autopercepção ou de sofrimento psíquico, nuance que se perde nesses casos.

Um terceiro padrão diz respeito à *desestruturação de enumerações de sintomas*. Diversas postagens do *corpus* original apresentam listas paralelas de sintomas (por exemplo, "depressão, ansiedade, auto estima e essas coisas"), e em vários casos a paráfrase reorganiza ou reduz essa enumeração, por vezes preservando apenas um subconjunto dos itens listados ou alterando a ordem de menção. Como o número e a combinação de sintomas relatados conjuntamente pode ser uma característica relevante para tarefas de classificação multirrótulo, essa reorganização nem sempre é semanticamente neutra.

Um quarto padrão é a *alteração do tipo de ato de fala* de algumas sentenças, isto é, a conversão de perguntas retóricas ou expressões de dúvida existencial (por exemplo, "Será que todo esse esforço vale a pena?") em sentenças declarativas. Perguntas retóricas são um recurso discursivo comum em relatos de sofrimento emocional, frequentemente utilizadas para expressar desesperança ou questionamento existencial sem expectativa literal de resposta; sua conversão sistemática em afirmações pode reduzir a expressividade do texto sintético em relação ao original.

Um quinto padrão observado foi o *tratamento inconsistente de placeholders de anonimização* presentes no *corpus* original (por exemplo, marcadores como `<Universidade19>`, utilizados para remover nomes próprios de instituições). Em alguns casos, esses marcadores foram preservados integralmente pela cadeia de tradução-paráfrase-retradução; em outros, foram parcialmente alterados ou teve sua posição na frase deslocada, o que pode gerar inconsistências caso o pipeline de anonimização dependa da localização exata desses marcadores para validação automática.

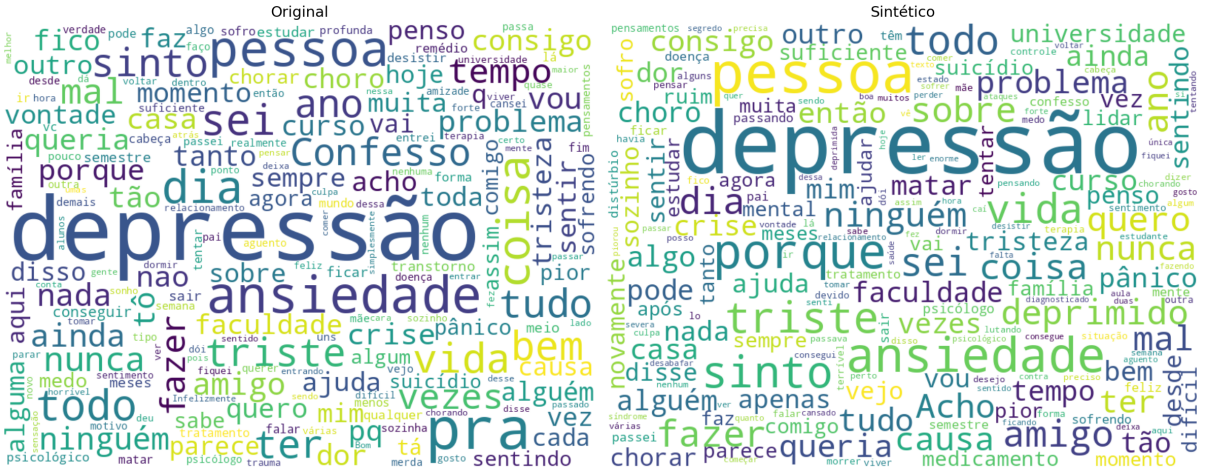
5.5.1.3 Análise das Nuvem de Palavras

Como forma complementar de avaliação qualitativa, foram geradas nuvens de palavras (*word clouds*) comparando a frequência lexical do *corpus* original e do *corpus* sintético, tanto para o conjunto de postagens relacionadas a sintomas depressivos quanto para o conjunto de controle (postagens sem indicação de sintomatologia depressiva). Essa visualização permite avaliar, em nível agregado e sem comprometer os dados reais, se a distribuição temática do *corpus* é preservada pelo pipeline, complementando a análise individual realizada até então.

Na comparação referente ao conjunto de postagens com sintomas depressivos (Figura 11), observa-se que o vocabulário temático central do *corpus* original é preservado de forma consistente no *corpus* sintético: termos como "depressão", "ansiedade", "tristeza", "pânico" e "crise" mantêm proeminência visual semelhante em ambas as nuvens, o que indica que a frequência relativa do núcleo lexical clínico do *corpus* não foi distorcida pelo pipeline em nível agregado. Este resultado é positivo, considerando que esse vocabulário é o principal sinal discriminativo para a tarefa de classificação de sintomas. A diferença mais perceptível entre as duas nuvens não está, portanto, na presença ou ausência dos

termos clínicos centrais, mas em marcadores de *oralidade e informalidade* característicos do discurso espontâneo em redes sociais. No *corpus* original, observam-se com destaque contrações e formas reduzidas típicas da escrita coloquial, como “pra” e “q”, além de verbos e expressões de fala direta (“acho”, “sei”, “confesso”). No *corpus* sintético, essas formas são sistematicamente normalizadas: “pra” não aparece entre os termos mais frequentes, sendo substituída por sua forma plena “para” (presente, porém com frequência reduzida, na nuvem do conjunto de controle), e abreviações como “q” desaparecem completamente. Esse padrão é coerente com a tendência de neutralização do registro coloquial já discutida na análise individual de exemplos, evidenciando-se aqui também em nível agregado.

Figura 11 – Comparação das nuvens de palavras do *corpus* original e do *corpus* sintético (**classe positiva**)



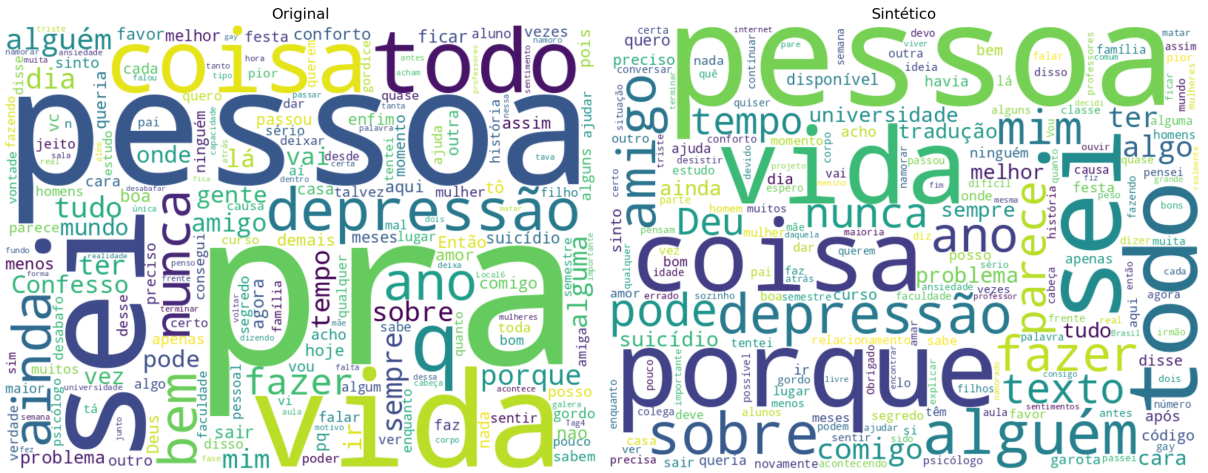
Um efeito particularmente interessante ocorre com o termo “Confesso”, presente com destaque na nuvem do *corpus* original e praticamente ausente na nuvem do *corpus* sintético. No *corpus* original, “confesso” funciona como marcador discursivo de abertura de relatos pessoais, sinalizando um enunciado de teor confessional, recorrente em postagens de desabafo em redes sociais. A ausência desse marcador no *corpus* sintético sugere que o pipeline de paráfrase tende a eliminar fórmulas de abertura discursiva desse tipo, convertendo o que era originalmente enunciado como uma confissão (“Confesso que estou com depressão...”) em uma afirmação direta e objetiva (“Estou com depressão...”). Embora o conteúdo proposicional central seja preservado nesses casos, a perda desse tipo de marcador discursivo reduz o caráter subjetivo e confessional do relato, que caracteriza o *corpus* original.

No conjunto de controle (Figura 12), a comparação entre as duas nuvens mostra que o vocabulário predominantemente genérico do *corpus* original (“pessoa”, “coisa”, “todo”, “vida”, “tempo”) é mantido no *corpus* sintético, com termos clínicos como “depressão” e “suicídio” permanecendo em proeminência reduzida em ambas as versões, condizente com a natureza desse subconjunto. A presença desses termos clínicos mesmo na classe de

controle, ainda que em menor proeminência, é esperada: todas as postagens do *corpus* Amive (inclusive aquelas das quais provêm as instâncias de controle) foram coletadas a partir de palavras-chave associadas à depressão e ideação suicida. As instâncias de controle correspondem a trechos dessas mesmas postagens para os quais nenhum sinal foi anotado pelos especialistas, o que pode ocorrer, por exemplo, quando o autor menciona depressão em referência a outra pessoa, oferece apoio a terceiros ou aborda o tema de forma contextual, sem que o trecho configure um relato pessoal de sofrimento. Essa distinção entre depressão como tema e como sintoma relatado em primeira pessoa é justamente o que torna a tarefa de classificação nesse domínio particularmente desafiadora.

A principal diferença qualitativa está, novamente, no registro discursivo: além da perda do marcador confessional "Confesso" e da normalização de contrações como "pra", nota-se uma distribuição lexical um pouco mais homogênea no *corpus* sintético, sem um termo isoladamente dominante como ocorre em "pessoa" no original. Isso sugere maior dispersão vocabular introduzida pela paráfrase nesse conjunto, possivelmente por se tratar de textos com menor densidade de vocabulário clínico e específico e, portanto, mais suscetíveis à reformulação livre pelo modelo.

Figura 12 – Comparação das nuvens de palavras do *corpus* original e do *corpus* sintético (**classe de controle**)



Em conjunto, a análise das nuvens de palavras sugere que o pipeline preserva adequadamente o vocabulário temático central de ambos os conjuntos, mas tende a uniformizar o estilo discursivo do *corpus*, eliminando marcadores de oralidade e de abertura confessional que caracterizam o gênero textual original. Esse achado complementa, em nível agregado, os padrões de neutralização de registro e de alteração de ato de fala já identificados na análise qualitativa de exemplos individuais.

5.5.2 Discussão Conjunta dos Resultados

Os resultados apresentados ao longo deste capítulo analisam e interpretam o que foi obtido até então, permitindo uma leitura conjunta sobre a qualidade do *corpus* sintético produzido para o sintoma de Tristeza/Humor depressivo.

Em primeiro lugar, as estatísticas do processo de geração indicam que o pipeline produziu, para a grande maioria das instâncias, paráfrases com similaridade semântica satisfatória em relação ao texto original: o BERTScore F1 médio de 75,90 na classe positiva e 74,66 no controle situa-se em uma faixa que, embora não represente equivalência semântica completa, é compatível com reformulações que preservam o conteúdo proposicional central ao mesmo tempo em que introduzem variação textual real. Isso corrobora com o valor obtido para o índice de Jaccard em torno de 54% a 55% entre original e paráfrase.

Essa combinação de similaridade semântica moderada-alta com diferença lexical superior à metade do vocabulário é, em certo sentido, o resultado desejado de um pipeline de geração de dados sintéticos: paráfrases idênticas ao original (BERTScore próximo de 100 e Jaccard próximo de 0) ofereceriam pouca proteção de privacidade e nenhuma diversidade textual nova, enquanto paráfrases muito dissimilares (BERTScore baixo) arriscariam descaracterizar o conteúdo clínico relevante.

Os 43 casos da classe positiva e os 23 da classe de controle que precisaram ser substituídos por apresentarem score exatamente 0 ou 100 confirmam, por outro lado, que esse equilíbrio não é automático: trata-se de uma minoria de instâncias, mas suficiente para exigir uma etapa adicional de verificação no pipeline, o que reforça a necessidade de validação manual ou semi-automática em qualquer aplicação futura desse método.

A avaliação intrínseca de diversidade lexical (Tabela 6 e Tabela 7) é consistente com esse panorama. O D-2 superior no *corpus* sintético em relação ao original, tanto na classe positiva (de 0,6575 para 0,7243) quanto na classe de controle (de 0,7489 para 0,7796), indica que, apesar da leve redução no tamanho absoluto do vocabulário, as paráfrases geradas combinam palavras de formas menos repetitivas do que os textos originais. Esse é um indício de que o pipeline não apenas substitui palavras isoladas, mas efetivamente reestrutura as sentenças. O BLEU invertido baixo em ambas as classes (0,1394 na positiva e 0,1218 no controle) reforça essa leitura: a baixa sobreposição de n-gramas com o texto de referência mostra que a reformulação não se limita a trocas pontuais de sinônimos, sendo coerente com o padrão observado nos exemplos qualitativos de inversão de estrutura causal e reorganização sintática (subseção 5.5.1). É digno de nota que esses indicadores de diversidade lexical foram sistematicamente mais favoráveis na classe de controle do que na positiva (vocabulário reduzido proporcionalmente mais, D-2 mais alto, BLEU invertido mais baixo), o que antecipa, em nível agregado, a discussão sobre a maior liberdade de reformulação observada nesses textos.

O comprimento dos textos, analisado na subseção 5.1.4, adiciona uma camada importante a essa leitura. A redução de aproximadamente 26% no número médio de tokens, observada tanto na classe positiva quanto na de controle, como discutida na Subseção indicada, não decorre de uma limitação de entrada do PEGASUS (uma vez que textos longos já são segmentados antes da paráfrase) mas da dificuldade do modelo em preservar integralmente o conteúdo proposicional de sentenças com múltiplas orações coordenadas e subordinadas, estrutura típica do gênero desabafo. Essa poda sistemática de conteúdo é uma hipótese plausível para parte dos ganhos de diversidade lexical relativa observados no D-2 e no BLEU invertido: um texto mais curto e sintaticamente mais simples tende, por construção, a repetir menos combinações de palavras e a compartilhar menos n-gramas com o original mais longo, ainda que isso não represente necessariamente uma reformulação mais rica do ponto de vista comunicativo. Essa ressalva é relevante para a interpretação dos resultados de diversidade lexical como um todo: parte do ganho métrico pode refletir simplificação estrutural, e não apenas reformulação lexical genuína.

Voltando-se à avaliação extrínseca (Tabela 4 e Tabela 5, Figura 6 e Figura 8), o cenário Sintético → Sintético confirma, de forma consistente nos quatro classificadores, que o *corpus* sintético constitui um conjunto de dados internamente coerente e até mais favorável à classificação do que o original, com ganhos de acurácia e precisão em todos os modelos e ganhos particularmente expressivos para o Naive Bayes (precisão de 0,668 para 0,731) e o Random Forest (precisão de 0,826 para 0,897). Esse resultado, de acordo com a discussão anterior sobre redução de comprimento e simplificação sintática, admite uma leitura cautelosa: textos sintéticos mais curtos e estruturalmente mais homogêneos tendem a produzir representações TF-IDF menos esparsas e fronteiras de decisão mais nítidas entre as classes, o que pode inflar artificialmente o desempenho observado nesse cenário específico, sem que isso represente, por si só, uma evidência de que o *corpus* sintético generaliza melhor para dados reais.

É justamente por essa razão que o cenário Sintético → Original (no qual um classificador treinado exclusivamente com paráfrases é avaliado sobre os textos reais de teste) constitui o teste mais rigoroso da utilidade prática do *corpus* sintético, simulando a situação de uso pretendida: pesquisadores que recebem apenas o *corpus* distribuível e o utilizam para treinar modelos que serão aplicados a dados reais no mundo, indisponíveis a eles. Os resultados nesse cenário são heterogêneos entre métricas e modelos, mas revelam um padrão sistemático: a precisão cai em todos os quatro classificadores em relação ao baseline original, com as maiores quedas absolutas observadas no SVM (de 0,854 para 0,763, uma perda de 0,091) e na Regressão Logística (de 0,836 para 0,787, uma perda de 0,049), enquanto a revocação se mantém estável ou melhora na maioria dos modelos (SVM (de 0,836 para 0,850) e Random Forest (estável, de 0,779 para 0,779)), com queda apenas marginal no Naive Bayes (de 0,921 para 0,907) e na Regressão Logística (de 0,800 para 0,793). O Random Forest destaca-se como o único modelo em que praticamente todas as

métricas no cenário Sintético → Original igualam ou superam o baseline original, inclusive a precisão (de 0,826 para 0,865), o que sugere que sua estratégia de particionamento hierárquico, ao agregar decisões de múltiplas árvores treinadas sobre subamostras distintas de atributos, é particularmente robusta à variação lexical introduzida pelas paráfrases (consistente com a justificativa apresentada na subseção 4.1.2.2 para sua inclusão no conjunto de modelos avaliados).

Esse deslocamento sistemático em direção a uma maior revocação (ou, no caso do Random Forest, a uma melhora geral) às custas da precisão é particularmente relevante no domínio de aplicação deste trabalho. Como discutido na subseção 2.4.1, a revocação é a métrica mais crítica em tarefas de triagem de saúde mental, em que deixar de identificar um caso real de sintoma depressivo (falso negativo) tende a ter consequências mais graves do que um alarme falso (falso positivo). Sob essa ótica, o fato de o *corpus* sintético tornar os classificadores mais sensíveis à classe positiva, mesmo quando treinados exclusivamente com paráfrases e avaliados sobre dados reais, é um resultado encorajador para o objetivo central deste trabalho: ainda que o *corpus* sintético não seja um substituto perfeito do original, a direção do erro introduzido (menos falsos negativos, mais falsos positivos) é a mais favorável dentre as duas possíveis para uma aplicação de triagem.

O painel complementar, Original → Sintético, no qual um classificador treinado em dados reais é avaliado sobre o *corpus* sintético, mostra o padrão inverso: a precisão permanece relativamente estável, mas a revocação cai de forma mais acentuada, com perdas de até 0,050 tanto no SVM quanto no Random Forest. Em conjunto, os dois cenários cruzados sugerem que a direção da transferência de conhecimento entre os domínios original e sintético não é simétrica: é mais fácil para um modelo treinado em dados sintéticos recuperar corretamente instâncias positivas reais do que para um modelo treinado em dados reais recuperar corretamente instâncias positivas sintéticas, possivelmente porque a poda sistemática de conteúdo e a simplificação sintática observadas no *corpus* sintético eliminam parte da variabilidade presente nos textos reais, tornando o vocabulário sintético, em algum sentido, um subconjunto mais restrito e mais previsível do vocabulário original.

De modo geral, a discussão aqui apresentada indica que o *corpus* sintético gerado constitui um substituto funcional, mas não perfeito, do *corpus* original: adequado a aplicações que toleram alguma perda de precisão em favor da sensibilidade na detecção de casos positivos, e cuja qualidade, embora majoritariamente satisfatória, não dispensa verificação adicional antes de sua distribuição ou uso em contextos clínicos de maior responsabilidade.

6 Conclusões

Este trabalho propôs e avaliou um pipeline de geração de dados sintéticos para o *corpus* Amive, voltado à detecção de sintomas de depressão a partir de postagens em redes sociais. O pipeline combina tradução automática via modelo de linguagem (PT→EN→PT), geração de paráfrases com o PEGASUS e seleção da melhor candidata por meio do BERTScore calculado com o BERTimbau, produzindo um *corpus* sintético com correspondência um-para-um em relação ao *corpus* original.

Do ponto de vista do objetivo central deste trabalho, isto é, viabilizar a distribuição púbica de um *corpus* que não pode ser disponibilizado em sua forma original devido a restrições éticas e de privacidade, os resultados obtidos são positivos. Pesquisadores que posteriormente receberem o *dataset* sintético poderão utilizá-lo para treinar classificadores variados e obter desempenhos comparáveis aos que obteriam com os dados originais, especialmente no que diz respeito à revocação (a métrica mais relevante para aplicações de triagem em saúde mental, em que minimizar falsos negativos é prioritário). A avaliação intrínseca corrobora essa conclusão: os valores de D-2 e BLEU invertido indicam que o *corpus* sintético introduz diversidade lexical genuína, preservando ao mesmo tempo grande parte do vocabulário clinicamente relevante, conforme observado na análise qualitativa das nuvens de palavras.

Ainda assim, os resultados também revelam que essa preservação não é uniforme: a precisão dos classificadores tende a cair de forma mais acentuada e consistente do que a revocação quando treinados exclusivamente com dados sintéticos e avaliados sobre dados reais, e a classe de controle sofre maior distorção lexical e sintática do que a classe positiva ao longo do pipeline. Esses resultados, discutidos em detalhe na seção 5.5, indicam que o *corpus* sintético constitui um substituto funcional, mas não perfeito, do *corpus* original, sendo mais adequado a aplicações que toleram alguma perda de precisão em favor da sensibilidade na detecção de casos positivos.

6.1 Limitações

Algumas limitações deste trabalho devem ser consideradas na interpretação dos resultados apresentados.

Em primeiro lugar, os experimentos de classificação foram conduzidos exclusivamente para o sintoma de Tristeza/Humor depressivo, o mais frequente do *corpus* Amive. Sintomas com um número menor de instâncias anotadas, como Agitação ou inquietação e Déficit de atenção ou memória, podem apresentar resultados distintos, possivelmente

inferiores, dado que conjuntos de treino menores tendem a ser mais sensíveis a variações lexicais introduzidas pelo processo de paráfrase.

Em segundo lugar, o processo de tradução automática pode comprometer expressões idiomáticas, gírias e abreviações típicas da linguagem informal de redes sociais. Como o pipeline traduz os textos para o inglês antes de gerar as paráfrases e retraduzi-los ao português, expressões coloquiais sem correspondência direta no idioma intermediário podem ser perdidas ou substituídas por formulações mais formais, alterando sutilmente o registro linguístico original dos textos.

Relacionada a essa limitação está a possibilidade de que parte dos ganhos de desempenho observados no cenário Sintético → Sintético decorra justamente dessa perda de nuances comunicativas: ao uniformizar o registro linguístico e reduzir peculiaridades únicas de cada indivíduo ao escrever, presentes em textos espontâneos, o *corpus* sintético pode se tornar artificialmente mais fácil de classificar, sem que isso represente necessariamente uma melhoria na capacidade de generalização para textos reais com a mesma variabilidade do *corpus* original.

Outra limitação importante diz respeito ao critério de seleção da melhor paráfrase. O BERTScore avalia a similaridade semântica de forma global, considerando o significado geral do texto. Isso implica que uma paráfrase pode obter pontuação elevada mesmo que tenha perdido ou suavizado um marcador clínico específico, desde que o conteúdo geral da frase permaneça semanticamente próximo ao original. Como discutido na Seção seção 5.5, o pipeline não possui qualquer mecanismo que reconheça ou priorize explicitamente esses marcadores, de modo que sua preservação observada nos resultados é incidental, e não uma garantia metodológica.

Por fim, a avaliação realizada neste trabalho concentrou-se em métricas quantitativas de diversidade lexical e desempenho de classificadores. Uma avaliação qualitativa mais aprofundada dos dados gerados - por exemplo, verificando sistematicamente se as paráfrases produzidas fazem sentido gramatical e semântico, e quantificando de forma mais granular o quão distintas elas de fato estão dos textos originais além das métricas agregadas aqui utilizadas - não foi conduzida de forma extensiva, e poderia revelar problemas específicos não capturados pelas métricas adotadas.

6.2 Trabalhos Futuros

A partir das limitações identificadas, algumas direções são sugeridas para a continuidade desta pesquisa.

Primeiramente, sugere-se aplicar o pipeline a todos os 18 sintomas anotados no *corpus* Amive, e não apenas ao sintoma de Tristeza/Humor depressivo, permitindo avaliar

se os padrões observados se mantêm em sintomas com menor número de instâncias e com outras nuances linguísticas como a maior presença de linguagem abstrata e figurada. Nesse mesmo sentido, seria relevante explorar a tarefa de classificação multirrótulo, considerando a co-ocorrência de múltiplos sintomas em uma mesma postagem, de forma mais próxima ao cenário clínico real.

Em segundo lugar, recomenda-se investigar o impacto de diferentes configurações dos modelos utilizados no pipeline de geração, como o tamanho do *batch* de tradução, o número de *beams* e de grupos de diversidade na geração de paráfrases pelo PEGASUS, parâmetros que neste trabalho foram fixados em valores únicos por restrições de tempo e custo computacional, baseando-se nas sugestões de configurações dos trabalhos relacionados pertinentes.

Também é proposto o ajuste fino (*fine-tuning*) do BERTimbau especificamente para o domínio de saúde mental, o que poderia melhorar a sensibilidade do BERTScore a marcadores clínicos relevantes, endereçando parcialmente a limitação discutida anteriormente sobre a natureza global da métrica de similaridade.

Quanto ao critério de desempate entre paráfrases com BERTScore igual, atualmente resolvido pela ordem de geração do PEGASUS, que corresponde à sequência de maior probabilidade acumulada na busca em feixe, sugere-se investigar critérios alternativos, como a priorização de paráfrases com maior variedade de vocabulário em relação às demais candidatas.

Por fim, propõe-se explorar algumas variações do pipeline de geração detalhadas a seguir. A primeira consiste em manter mais de uma sentença sintética por instância original, em vez de selecionar apenas a melhor paráfrase, configurando uma estratégia de aumento de dados propriamente dita, na qual todas as traduções e paráfrases acima de um limiar de qualidade seriam preservadas, ampliando o *corpus* sintético além da correspondência um-para-um adotada neste trabalho.

A segunda consiste em avaliar uma versão simplificada do pipeline, sem a etapa de paráfrase via PEGASUS, utilizando apenas o processo de *back-translation* (tradução e retradução) como mecanismo de geração. Essa alternativa não foi adotada como abordagem principal neste trabalho por uma consideração prática: sistemas de tradução automática neural atuais, como o modelo de linguagem utilizado no pipeline proposto, tendem a produzir traduções de alta qualidade e consistência, o que reduziria significativamente a variação textual introduzida apenas pelo processo de ida e volta entre idiomas. Sem a etapa intermediária de paráfrase, é provável que grande parte das instâncias retraduzidas resultasse em textos muito próximos, ou até idênticos, aos originais, comprometendo o objetivo de gerar diversidade lexical suficiente para um *corpus* sintético distribuível. Ainda assim, uma avaliação sistemática dessa hipótese permitiria isolar a contribuição específica de cada componente do pipeline para a qualidade e a diversidade do *corpus* sintético final.

Como alternativa às variações descritas, também seria interessante investigar uma versão do pipeline sem qualquer etapa de tradução, empregando diretamente um modelo de paráfrase pré-treinado em português, como o PTT5-Paraphraser, tratado na Tabela 1. Essa abordagem eliminaria a dependência de tradução automática e, consequentemente, os riscos de distorção semântica associados a ela, ainda que ao custo de empregar um modelo de paráfrase potencialmente menos robusto e treinado em um *corpus* de menor escala do que os disponíveis para o inglês.

Considerações éticas

Os dados usados nesta pesquisa são postagens anônimas coletadas automaticamente de uma plataforma pública de rede social. Essas postagens integram um *corpus* disponibilizado gratuitamente a pesquisadores mediante solicitação. Os textos não contêm informações que permitam identificar os autores das postagens, protegendo assim a sua privacidade e minimizando potenciais danos. Os pesquisadores foram informados de que o conteúdo das postagens poderia incluir material sensível relacionado à saúde mental. Todas as análises foram realizadas utilizando métodos estabelecidos e amplamente reconhecidos.

Declaração do uso de Inteligência Artificial Generativa

O desenvolvimento deste trabalho contou com o auxílio de ferramentas de inteligência artificial em diferentes etapas. O modelo Claude (Anthropic)[1], acessado via interface claude.ai, foi utilizado como apoio na escrita e revisão dos textos acadêmicos, sempre com revisão, adaptação e validação pela autora. Na etapa de desenvolvimento, o mesmo modelo auxiliou na revisão do código Python dos pipelines de geração de paráfrases e classificação, bem como na geração das figuras de visualização dos resultados (gráficos de barras, mapa de calor e gráfico radar).

[1] <https://www.anthropic.com/claude>

APÊNDICE A – Diagramas detalhados do *pipeline* de geração de dados sintéticos

Este apêndice apresenta representações visuais (fluxogramas) de cada etapa do *pipeline* de geração de dados sintéticos descrito no Capítulo 4, incluindo parâmetros, critérios de decisão e tratamento de casos especiais.

Figura 13 – Fluxograma do processo de tradução automática via LLM.

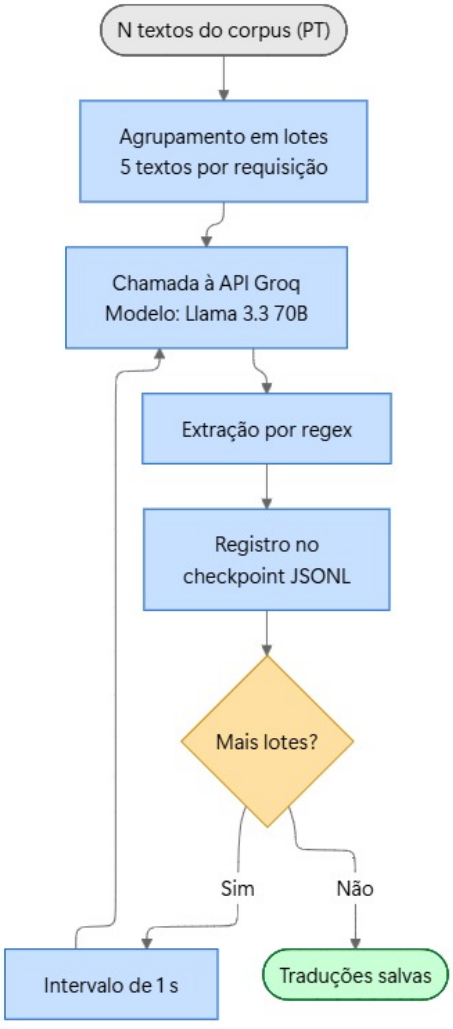


Figura 14 – Fluxograma do processo de geração de paráfrases com PEGASUS e *diverse beam search*.

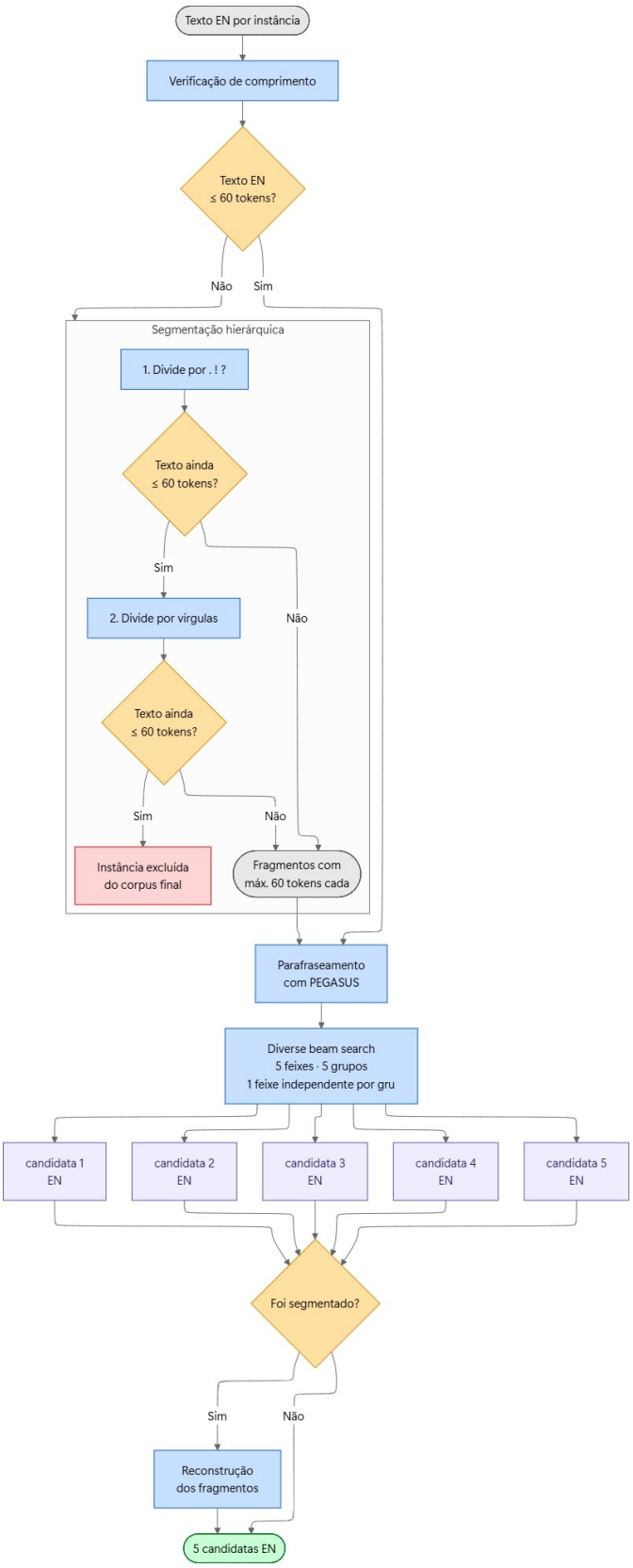
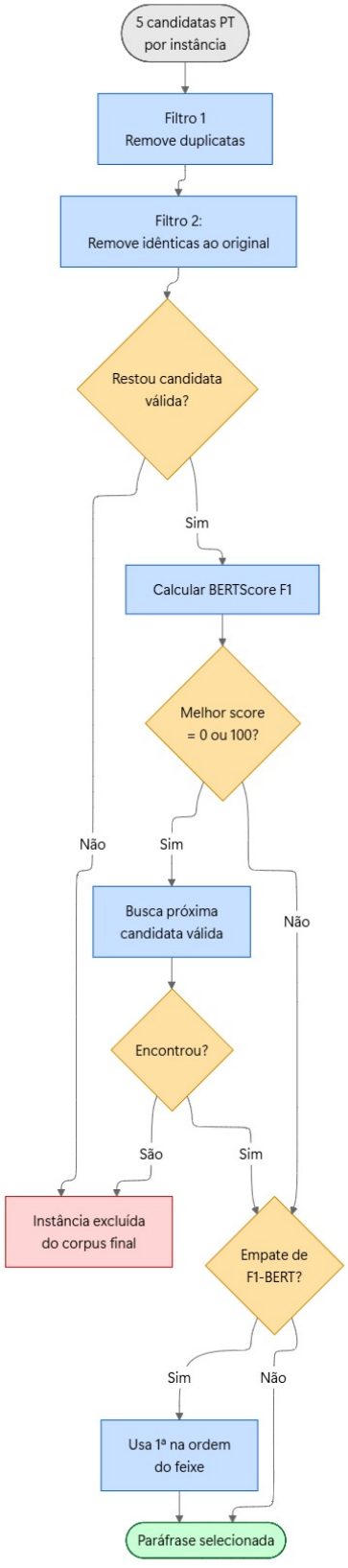


Figura 15 – Fluxo de decisão para filtragem e seleção da paráfrase via BERTScore.



Referências

AI@Meta. *Llama 3.3*. 2024. Disponível em: <https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>. Citado 2 vezes nas páginas 45 e 52.

ALHAMED, F. et al. Monitoring depression severity and symptoms in user-generated content: An annotation scheme and guidelines. In: CLERCQ, O. D. et al. (Ed.). *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. Bangkok, Thailand: Association for Computational Linguistics, 2024. p. 227–233. Disponível em: <https://aclanthology.org/2024.wassa-1.18/>. Citado na página 37.

ALHAMED, F.; IVE, J.; SPECIA, L. Classifying social media users before and after depression diagnosis via their language usage: A dataset and study. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. [S.l.: s.n.], 2024. p. 3250–3260. Citado na página 21.

BRAINS – Brazilian AI Networks. *Os Transformers Ilustrados*. 2023. Disponível em: <https://brains.dev/2023/os-transformers-ilustrados/>. Acesso em: 1 jul. 2026. Citado na página 28.

Brasil. *Lei nº 13.709, de 14 de agosto de 2018 — Lei Geral de Proteção de Dados Pessoais (LGPD)*. 2018. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>. Citado na página 23.

CASTILHO, S.; CASELI, H. d. M. Tradução automática: Abordagens e avaliação. In: CASELI, H. M.; NUNES, M. G. V. (Ed.). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. 4. ed. BPLN, 2026. v. 3, book chapter 4. ISBN 978-65-02-00167-7. Disponível em: <https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/aplicacoes-e-dominios/cap-mt/cap-mt.html>. Citado na página 29.

CHIM, J.; IVE, J.; LIAKATA, M. Evaluating synthetic data generation from user generated text. *Computational Linguistics*, MIT Press, Cambridge, MA, v. 51, n. 1, p. 191–233, mar. 2025. Disponível em: <https://aclanthology.org/2025.cl-1.6/>. Citado 2 vezes nas páginas 41 e 42.

CHOUDHURY, M. D. et al. Predicting depression via social media. In: *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)*. [S.l.: s.n.], 2013. p. 128–137. Citado na página 21.

European Parliament and Council. *Regulation (EU) 2016/679 — General Data Protection Regulation*. [S.l.], 2016. Disponível em: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>. Citado na página 23.

HONNIBAL, M. et al. *spaCy: Industrial-strength Natural Language Processing in Python*. 2020. Citado na página 52.

IGAMBERDIEV, T.; HABERNAL, I. Dp-bart for privatized text rewriting under local differential privacy. In: *Findings of the Association for Computational Linguistics: ACL*

2023. Association for Computational Linguistics, 2023. p. 13914–13934. Disponível em: <http://dx.doi.org/10.18653/v1/2023.findings-acl.874>. Citado 2 vezes nas páginas 21 e 38.

IPPOLITO, D. et al. Comparison of diverse decoding methods from conditional language models. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2019. p. 3752–3762. Citado na página 32.

JOHANSSON, K. et al. Meta4xnli-ptbr: Brazilian portuguese extension of meta4xnli corpus. In: PIPERIDIS, S. et al. (Ed.). *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*. Palma, Mallorca, Spain: European Language Resources Association (ELRA), 2026. p. 1668–1676. Citado 3 vezes nas páginas 40, 41 e 46.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. [S.l.]: Cambridge University Press, 2008. Citado 3 vezes nas páginas 30, 31 e 59.

MAWATWALA, M. *Word Embeddings — LLM Concept*. 2023. Disponível em: <https://medium.com/@mawatwalmanish1997/evolution-of-word-embeddings-48d2449e933d>. Acesso em: 1 jul. 2026. Citado na página 26.

MENDES, A. R.; CASELI, H. d. M. 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)*. Torino, Italia: ELRA and ICCL, 2024. p. 8594–8604. Citado na página 50.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. Citado na página 26.

MOHAMMADI, M.; AMIN, M. R.; TAVAKOLI, S. Boosting sentiment analysis in Persian through a GAN-based synthetic data augmentation method. In: EL-HAJ, M. et al. (Ed.). *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*. Abu Dhabi, UAE: Association for Computational Linguistics, 2025. p. 54–63. Disponível em: <https://aclanthology.org/2025.abjadnlp-1.7/>. Citado 3 vezes nas páginas 21, 39 e 41.

PAES, A.; VIANNA, D.; RODRIGUES, J. Modelos de linguagem. In: CASELI, H. M.; NUNES, M. G. V. (Ed.). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. 4. ed. BPLN, 2026. v. 2, book chapter 1. ISBN 978-65-02-00158-5. Disponível em: <https://brasileiraspln.ufscar.br/livro-pln-4ed-vol2/modelos/cap-modelos-linguagem/cap-modelos-linguagem.html>. Citado na página 25.

PAGANO, A. et al. Pln na saúde. In: CASELI, H. M.; NUNES, M. G. V. (Ed.). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. 4. ed. BPLN, 2026. v. 3, book chapter 9. ISBN 978-65-02-00167-7. Disponível em: <https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3/aplicacoes-e-dominios/cap-saude/cap-saude.html>. Citado 2 vezes nas páginas 21 e 23.

PAPINENI, K. et al. BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2002. p. 311–318. Citado na página 33.

PARABONI, I.; CASELI, H. d. M. Pln na saúde mental: Detecção de transtornos de saúde mental a partir de texto. In: CASELI, H. M.; NUNES, M. G. V. (Ed.). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. 3. ed. BPLN, 2024. book chapter 29. ISBN 978-65-01-20581-6. Disponível em: <https://brasileiraspln.com/livro-pln/3a-edicao/parte-dominios/cap-saude-mental/cap-saude-mental.html>. Citado 3 vezes nas páginas 21, 37 e 50.

PATSAKIS, C.; LYKOUSAS, N. Man vs the machine in the struggle for effective text anonymisation in the age of large language models. *Scientific Reports*, Nature Publishing Group, v. 13, n. 1, p. 16026, 2023. Citado na página 24.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 52.

PELLICER, L. F. A. O. et al. PTT5-Paraphraser: Diversity and meaning fidelity in automatic Portuguese paraphrasing. In: *Proceedings of the 15th International Conference on Computational Processing of the Portuguese Language (PROPOR 2022)*. [S.l.]: Springer, 2022. p. 299–309. Citado 3 vezes nas páginas 40, 41 e 42.

PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Disponível em: <https://aclanthology.org/D14-1162>. Citado na página 26.

RIJSBERGEN, C. J. V. *Information Retrieval*. 2nd. ed. London: Butterworths, 1979. Citado na página 32.

SANTOS, W. R. dos; OLIVEIRA, R. L. de; PARABONI, I. Setembrobr: a social media corpus for depression and anxiety disorder prediction. *Language Resources and Evaluation*, v. 58, n. 1, p. 273–300, 2024. Citado na página 21.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: Pretrained bert models for brazilian portuguese. In: _____. [S.l.: s.n.], 2020. p. 403–417. ISBN 978-3-030-61376-1. Citado na página 47.

STAAB, R. et al. Beyond memorization: Violating privacy via inference with large language models. In: *The Twelfth International Conference on Learning Representations (ICLR)*. [S.l.: s.n.], 2024. Citado na página 24.

SWEENEY, L. Simple demographics often identify people uniquely. *Carnegie Mellon University, Data Privacy Working Paper*, v. 3, 2000. Disponível em: <https://dataprivacylab.org/projects/identifiability/paper1.pdf>. Citado na página 24.

The Pandas Development Team. *pandas-dev/pandas: Pandas*. 2020. Disponível em: <https://pandas.pydata.org>. Citado na página 52.

VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017. Citado na página 24.

VIJAYAKUMAR, A. K. et al. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016. Citado 2 vezes nas páginas 46 e 53.

WAY, A.; FORCADA, M. L. Editors' foreword to the invited issue on smt and nmt. *Machine Translation*, Springer, v. 32, n. 3, p. 191–194, 2018. ISSN 09226567, 15730573. Disponível em: <http://www.jstor.org/stable/44987868>. Citado 2 vezes nas páginas 29 e 30.

WILKENS, R. et al. The visible and the latent linguistic clues of mental health in Brazilian Portuguese textual posts. In: SOUZA, M. et al. (Ed.). *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 2*. Salvador, Brazil: Association for Computational Linguistics, 2026. p. 135–147. ISBN 979-8-89176-387-6. Disponível em: <https://aclanthology.org/2026.propor-2.21/>. Citado 3 vezes nas páginas 15, 23 e 51.

WOLF, T. et al. Transformers: State-of-the-art natural language processing. In: *Proceedings of EMNLP 2020: System Demonstrations*. [S.l.: s.n.], 2020. p. 38–45. Citado na página 52.

ZHANG, F. et al. Clear up confusion: Iterative differential generation for fine-grained intent detection with contrastive feedback. In: RAMBOW, O. et al. (Ed.). *Proceedings of the 31st International Conference on Computational Linguistics*. Abu Dhabi, UAE: Association for Computational Linguistics, 2025. p. 2207–2221. Disponível em: <https://aclanthology.org/2025.coling-main.151/>. Citado 5 vezes nas páginas 21, 32, 33, 39 e 41.

ZHANG, J. et al. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In: III, H. D.; SINGH, A. (Ed.). *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020. (Proceedings of Machine Learning Research, v. 119), p. 11328–11339. Disponível em: <https://proceedings.mlr.press/v119/zhang20ae.html>. Citado 2 vezes nas páginas 22 e 46.

ZHANG, T. et al. *BERTScore: Evaluating Text Generation with BERT*. 2020. Disponível em: <https://arxiv.org/abs/1904.09675>. Citado 3 vezes nas páginas 22, 33 e 52.

ZIMMER, M. "but the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology*, v. 12, n. 4, p. 313–325, 2010. Citado na página 23.