UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

EDUARDO SANTIAGO MUNIZ FILHO

VALIDAÇÃO DE TÉCNICAS DE DATA AUGMENTATION EM CORPUS TEXTUAIS SOBRE DIABETES

PATO BRANCO

2023

EDUARDO SANTIAGO MUNIZ FILHO

VALIDAÇÃO DE TÉCNICAS DE DATA AUGMENTATION EM CORPUS TEXTUAIS SOBRE DIABETES

VALIDATION OF DATA AUGMENTATION TECHNIQUES IN CORPUS TEXTUALS ON DIABETES

Proposta de Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Computação do Curso de Bacharelado em Engenharia de Computação da Universidade Tecnológica Federal do Paraná.

Orientador: Profª. Drª Eliane Maria De Bortoli Fávero

PATO BRANCO

2023

EDUARDO SANTIAGO MUNIZ FILHO

VALIDAÇÃO DE TÉCNICAS DE DATA AUGMENTATION EM CORPUS TEXTUAIS SOBRE DIABETES

Proposta de Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Computação do Curso de Bacharelado em Engenharia de Computação da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 29/11/2023

Eliane Maria de Bortoli Fávero
Doutora em Ciência da Computação
Universidade Tecnológica Federal do Paraná

Rúbia Eliza de Oliveira Schultz Ascari
Doutora em Ciência da Computação
Universidade Tecnológica Federal do Paraná

Dalcimar Casanova
Doutor em Física Computacional
Universidade Tecnológica Federal do Paraná

PATO BRANCO

2023

AGRADECIMENTOS

Agradeço a Deus por todas as oportunidades que tive e por abençoar os momentos da minha vida com pessoas maravilhosas.

Sou grato a minha orientadora Profª. Drª Eliane Maria De Bortoli Fávero, pela sabedoria e todo suporte que me proporcionou nesta trajetória.

Gostaria de enfatizar também, o reconhecimento à minha mãe, pois nos momentos mais difíceis permaneceu ao meu lado e me encorajou a seguir em frente.

Aos professores, agradeço por todas as contribuições e ensinamentos no processo de formação acadêmica.

A Secretaria do Curso, pela cooperação.

Enfim, agradeço a todos que por algum motivo contribuíram na jornada acadêmica e pessoal que segui até a realização desta pesquisa.

RESUMO

A diabetes é uma doença crônica que afeta milhões de pessoas no mundo. Com o auxílio de modelos computacionais é possível auxilíar no controle e até prevenção da diabetes. Esses modelos podem ser baseados em métodos de aprendizado de máquina, que objetivem o reconhecimento de padrões de comportamento, por exemplo, em conjuntos de dados textuais de redes sociais, como o Twitter. Entretanto, a eficiência desses modelos inteligentes, está relacionada com o volume de dados disponíveis, bem como sua qualidade. Nesse contexto, o presente estudo propõe a validação das técnicas de aumento de dados em um *corpus* sobre diabetes proveniente do Twitter, com o objetivo de estabelecer uma fonte confiável e de qualidade para futuras pesquisas envolvendo processamento de linguagem natural aplicada à essa área específica da saúde. Os resultados demonstram que em algumas amostras do corpus o significado das sentenças foi alterado após o aumento dos dados, porém essa alteração não foi significativa para afetar a eficiência, garantindo ao corpus aumentado um desempenho melhor em modelos de classificação de texto.

Palavras-chave: diabetes; dados; data augmentation; métodos; aprendizado de máquina.

ABSTRACT

Diabetes is a chronic disease that affects millions of people worldwide. With the assistance of computational models, it is possible to aid in the control and even prevention of diabetes. These models can be based on machine learning methods that aim to recognize patterns of behavior, for example, in textual datasets from social networks, such as Twitter. However, the efficiency of these intelligent models is linked to the volume and quality of available data. In this context, the study proposes to validate data augmentation techniques in a corpus related to diabetes derived from Twitter, with the aim of establishing a reliable and high-quality source for future research in natural language processing applied to this specific area of health. The results demonstrate that in some samples of the corpus, the meaning of sentences was altered after data augmentation; however, this change was not significant enough to affect efficiency, ensuring that the augmented corpus performs better in text classification models.

Keywords: diabetes; data; data augmentation; methods; machine learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

Siglas

API	*Application Programming Interface*
BERT	*Bidirectional Encoder Representations from Transformers*
C-MLM	*Conditional Masked Language Model*
CSV	*Comma-Separated Values*
DA	*Data Augmentation*
DL	*Deep Learning*
EDA	*Easy Data Augmentation*
IDF	*International Diabetes Federation*
ML	*Machine Learning*
NB	*Naïve Bayes*
NLG	*Natural Language Generation*
NLP	*Natural Language Processing*
NLTK	*Natural Language Toolkit*
NLU	*Natural Language Understanding*
RF	*Random Forest*
RoBERTa	*Robustly Optimized BERT Approach*
SVM	*Support Vector Machine*
t-SNE	*t-Distributed Stochastic Neighbor Embedding*

SUMÁRIO

1 INTRODUÇÃO

A diabetes é uma doença crônica que ocorre quando o pâncreas não é mais capaz de produzir insulina, ou quando o corpo não consegue fazer bom uso da insulina que produz. Atualmente, de acordo com dados da *International Diabetes Federation* (IDF), estima-se que o número de pessoas com diabetes no mundo seja de 643 milhões em 2030, e 783 milhões até 2045 (Federation, 2021).

O aumento da prevalência[1] do diabetes está associado a diversos fatores, como transição epidemiológica, transição nutricional, aumento do sedentarismo, excesso de peso, crescimento e envelhecimento populacional e, também, à maior sobrevida dos indivíduos com diabetes (Diabetes, 2019).

Avanços da computação na área da saúde, principalmente em pesquisas envolvendo *Machine Learning* (ML) e *Deep Learning* (DL), possuem expressivas contribuições no estudo de padrões e comportamentos associados à condição clínica do diabetes (Amuthadevi; Umamaheswari; Belinda, 2018). Um estudo proposto por (Kumar; Umatejaswi, 2017), utiliza algoritmos de classificação *Random Forest*, *Naive Bayes* e *Support Vector Machine* (SVM) para identificar diabetes por meio de técnicas de mineração de dados. Outro estudo apresenta uma abordagem para *Diabetic Retinopathy*[2], uma das principais causas de problemas relacionados à visão em pessoas diabéticas (Pal; Poray; Sen, 2017).

De acordo com IDF, diferentes metodologias, aplicadas de acordo com o país, em seus dados brutos provenientes de estudos específicos, torna desafiador estimar o impacto global do diabetes. A incapacidade de fornecer cobertura abrangente de prevalência de diabetes também se deve à pura falta de dados sobre o tema ao redor do mundo (Federation, 2021). Ao analisar o espectro de estudos que utilizam dados textuais relacionados ao diabetes, esse cenário é ainda mais desafiador.

Na maioria das vezes, dados textuais relacionados ao diabetes, sejam publicações médicas, comentários em mídias sociais, registros médicos eletrônicos, manuais informativos, dentre outros, tendem a ser minimamente representadas como dados narrativos, em sua forma nativa e não podem ser analisados com técnicas analíticas tradicionais (Turchin; Builes, 2021). Esse cenário revela a pouca disponibilidade de dados textuais, a fim de fomentar pesquisas na área.

Natural Language Processing (NLP) é uma área da ciência da computação, na qual ML e a linguística computacional são amplamente utilizados. A máquina aprende a sintaxe e o significado da linguagem humana, processa e retorna um resultado esperado (Jain; Kulkarni; Shah, 2018). As ferramentas de NLP e esquemas de anotações de corpus[3] linguísticos, supor-

[1] Número total de casos existentes numa determinada população e num determinado momento temporal.

[2] Uma complicação da diabetes causada por danos aos vasos sanguíneos na retina.

[3] Coleção de textos autênticos, legíveis por máquina, que são amostrados para serem representativos de uma determinada língua natural ou variedade linguística. O plural de corpus é dado pela palavra corpora.(McEnery; Xiao; Tono, 2006)

tam a identificação automática de uma série de propriedades da linguagem, incluindo aspectos lexicais, sintáticos, semânticos e pragmáticos do sistema linguístico (Garside; Leech; McEnery, 1997).

O alto volume de pesquisas com foco na extração de informações do paciente em registros eletrônicos de saúde levou a um aumento na demanda por corpora anotados, sendo este um recurso valioso tanto para o desenvolvimento quanto para a avaliação de algoritmos de NLP (Oliveira *et al.*, 2022). Entretanto, com o aumento na demanda por corpora na área da saúde, ainda existe uma dificuldade considerável para encontrar corpus volumosos de textos sobre diabetes.

Data Augmentation (DA) é uma prática amplamente aceita para melhorar o treinamento de modelos de aprendizado de máquina e é especialmente útil quando os conjuntos de dados são limitados em tamanho. Ela desempenha um papel fundamental na redução do problema de *overfitting*[4] e no aumento do desempenho geral do modelo (LeCun; Bengio; Hinton, 2015). Segundo os autores, tratam-se de técnicas que envolvem a geração de novos exemplos de treinamento por meio da aplicação de transformações ou manipulações nos dados originais. Essa técnica é usada para expandir o conjunto de dados de treinamento, tornando-o mais diversificado e, assim, melhorando o desempenho dos modelos de aprendizado de máquina.

As mídias sociais desempenham um papel significativo na vida moderna, proporcionando uma plataforma para comunicação, compartilhamento de informações, interação social e expressão pessoal. Os dados textuais coletados nas mídias sociais são extremamente valiosos e podem representar uma oportunidade para a análise e compreensão de problemas comuns relacionados à saúde (Karami *et al.*, 2018).

O Twitter, recentemente nomeado como X, é um serviço de rede social online que permite aos usuários participar de conversas em tempo real e acompanhar atualizações por meio de um *feed* de notícias. Devido a esta mudança na denominação ocorrer durante o desenvolvimento deste trabalho, o nome Twitter será mantido. Os dados textuais provenientes do Twitter podem ser utilizados em diversas aplicações de temas relacionados à saúde (ex. gripe (Culotta, 2010), diabetes (Harris *et al.*, 2013) e obesidade (Dahl; Hales; Turner-McGrievy, 2016)).

Sendo assim, considerando a necessidade existente por corpus de dados textuais, volumosos o suficiente para desenvolver pesquisas sobre diabetes, o problema de pesquisa que se apresenta é: **Como provar que técnicas de DA aplicadas a dados textuais sobre diabetes obtidos do Twitter, possibilitam a geração de um corpus consistente e confiável sobre diabetes, sem deturpar os dados?**

Portanto, o presente trabalho busca validar o desempenho das técnicas de DA em corpus contendo *tweets* sobre diabetes, objetivando aumentar o volume dos dados de forma con-

[4] Overfitting ocorre quando o modelo treinado se ajusta muito bem a um conjunto específico de observações de treinamento, mas não prediz bem os resultados de observações não usadas no treinamento (Dean, 2014).

fiável e com qualidade, a fim de promover futuras pesquisas envolvendo NLP aplicada a essa área específica da saúde.

1.1 Objetivos

1.1.1 Objetivo geral

Validar o uso de técnicas de DA para corpus textuais sobre diabetes mantendo a qualidade dos dados originais.

1.1.2 Objetivos específicos

- Obter um corpus com *tweets* sobre diabetes, na língua inglesa, que servirá de base para a aplicação de técnicas de DA.
- Implementar técnicas de DA;
- Validar técnicas de DA;
- Analisar a *accuracy* obtida pelas técnicas de validação;
- Indicar quais técnicas de DA apresentaram melhor desempenho.

1.2 Justificativa

A contribuição de aplicações envolvendo ML e DL na área da saúde enfatiza a importância de estudos para auxiliar no tratamento de doenças, como por exemplo, a diabetes. No entanto, ainda existe uma dificuldade considerável na obtenção de dados textuais de qualidade.

Apesar dos desafios, muitas técnicas de DA para NLP têm sido propostas, desde manipulações baseadas em regras (Zhang; Zhao; LeCun, 2015) até abordagens generativas mais complexas (Liu *et al.*, 2020). É esperado que uma abordagem eficaz de DA facilite a implementação e melhore o desempenho do modelo. Além disso, a distribuição dos dados aumentados não deve ser muito semelhante nem muito diferente dos dados originais (Feng *et al.*, 2021).

Portanto, validar o uso de técnicas de DA para a geração de corpus textuais sobre diabetes, oferece uma alternativa viável para aumentar a quantidade dos dados.

1.3 Estrutura do trabalho

- **Capítulo 1**: neste capítudo são apresentados de forma introdutória a temática da proposta, os objetivos e as motivações do trabalho.

- **Capítulo 2**: capítulo em que são abordados os conceitos e métodos de DA relacionados a NLP.
- **Capítulo 3**: este capítulo apresenta os trabalhos relacionados com a proposta que utilizam métodos de DA para expansão de corpus.
- **Capítulo 4**: capítulo contendo os materiais e métodos para o desenvolvimento do trabalho.
- **Capítulo 5**: neste capítulo são apresentados os resultados dos experimentos com as técnicas de DA.
- **Capítulo 6**: este capítulo apresenta a conclusão do desenvolvimento do trabalho.

2 REFERENCIAL TEÓRICO

2.1 *Machine Learning*

ML é um subcampo de estudo da ciência da computação direcionado ao aprendizado automático dos computadores. Seu desenvolvimento tem origem na área de reconhecimento de padrões e inteligencia artificial, com destaque para pesquisas em processamento de linguagem natural, visão computacional, reconhecimento de padrões, computação cognitiva e representação do conhecimento (N; Gupta, 2020a).

Um algoritmo de ML é um algoritmo capaz de aprender com os dados (Goodfellow; Bengio; Courville, 2016). É dito que um programa de computador aprende com a experiência *E* em relação a alguma classe de tarefas *T* e medida de desempenho *P*, se seu desempenho nas tarefas *T*, medido por *P*, melhora com a experiência *E* (Mitchell, 1997).

As tarefas são geralmente descritas em termos de como o sistema de ML deve processar uma coleção de recursos medidos quantitativamente. De acordo com a tarefa executada, é importante projetar uma medida de desempenho para avaliar o algoritmo. Conforme o tipo de experiência adquirida durante o processo de aprendizado os algoritmos de ML podem ser amplamente categorizados como não supervisionados ou supervisionados (Goodfellow; Bengio; Courville, 2016). Além desses, os algoritmos de aprendizado semi-supervisionado e de aprendizagem por reforço possuem grande relevância.

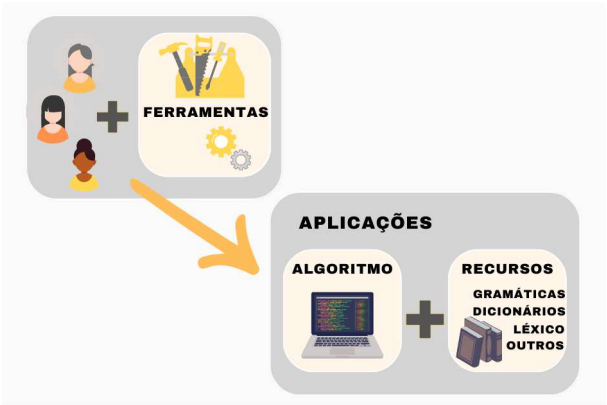
Os algoritmos de aprendizado supervisionado exploram os dados de treinamento e produzem uma função completa que pode ser utilizada para mapear novas instâncias. Desta forma, o algoritmo de aprendizagem será capaz de analisar e generalizar corretamente os rótulos da classe a partir de instâncias não observadas. Algoritmos de aprendizagem não supervisionada exploram um conjunto de dados para descobrir estruturas ocultas não rotulados ou inferir um modelo que tenha a densidade de probabilidade dos dados de entrada (N; Gupta, 2020a). O objetivo do aprendizado semi-supervisionado é classificar os dados não rotulados com base nos dados rotulados. Já o aprendizado por reforço, possibilita a aprendizagem a partir do *feedback* recebido por meio de interações com o ambiente externo (N; Gupta, 2020b).

2.2 *Natural Language Processing*

NLP é um campo de pesquisa que objetiva investigar e propor métodos e sistemas de processamento computacional da linguagem humana. Existem duas grandes subáreas da NLP. A primeira, *Natural Language Understanding* (NLU) visa a análise e a interpretação da língua, enquanto a segunda, *Natural Language Generation* (NLG) concentra-se na geração de linguagem natural (Caseli; Nunes, 2023).

No contexto de NLP as ferramentas auxiliam na construção de uma aplicação, seja um sistema computacional ou um aplicativo. As aplicações fornecem um resultado ao usuário com uma entrada ou saída em linguagem natural. As aplicações fazem uso de ferramentas ou conjuntos de ferramentas, conhecidos como *toolkits*. Além disso, os recursos fornecem informações linguísticas necessárias para que as aplicações consigam processar a língua da maneira adequada (Caseli; Nunes, 2023). Um exemplo dessa relação entre aplicações, recursos e ferramentas pode ser visto na Figura 1.

Figura 1 – Relacionamento entre aplicações, recursos e ferramentas.



Fonte: Adaptado de (Caseli; Nunes, 2023).

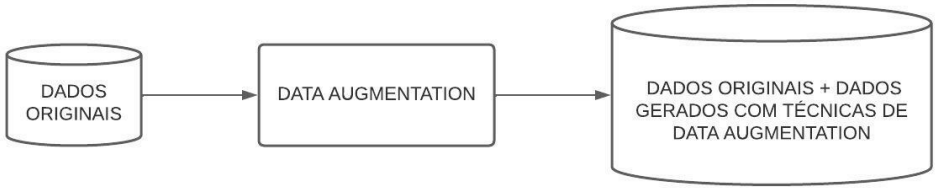
São definidos três paradigmas relacionados à NLP. O paradigma simbólico define que todo o conhecimento sobre a língua é representado de forma explícita em formalismos, como léxicos, regras, linguagens lógicas, e outras formas compreensíveis ao ser humano. No paradigma estatístico grandes conjuntos de textos são usados como fonte de conhecimento para “ensinar” as máquinas. Similar ao paradigma estatístico, o paradigma neural também se baseia em grandes volumes de dados para aprender um modelo, entretanto, a forma de aprendizado envolve múltiplas camadas de unidades de processamento para reconhecer os padrões recorrentes. Outras abordagens utilizam paradigmas híbridos, que combinam principalmente o paradigma simbólico com um dos demais (Caseli; Nunes, 2023).

2.3 *Data Augmentation*

O termo DA, refere-se a métodos usados para aumentar a quantidade de dados com base em diferentes técnicas de modificações desses dados, as quais permitem expandir um conjunto de dados original. De acordo com (Feng *et al.*, 2021), o aumento no volume de dados com uso de DA atua como um regularizador e ajuda a evitar o *overfitting* no treinamento de modelos de aprendizado de máquina. A Figura 2 mostra uma representação simplificada de DA.

Na computação visual, comumente são utilizadas técnicas de DA em imagens com transformações simples, como deslocamento horizontal, aumentos no espaço de cores e corte ale-

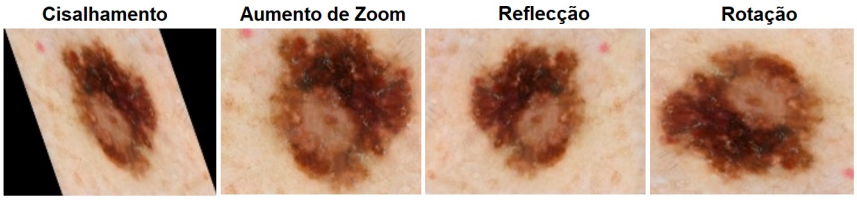
Figura 2 – Diagrama simplificado de DA.



Fonte: Autoria própria (2023).

atório para formar componentes do modelo de treinamento. A aplicação das técnicas de DA tem demonstrado um desempenho notável em diversas tarefas na computação visual (Shorten; Khoshgoftaar, 2019). A Figura 3 mostra um exemplo de DA em imagens para a classificação de melanoma.

Figura 3 – Transformações em imagens de melanoma utilizando DA.



Fonte: Adaptado de (Mikołajczyk; Grochowski, 2018).

Em NLP o espaço de dados de entrada é discreto, e gerar exemplos com aumento de dados eficazes que capturem as invariâncias desejadas, é menos óbvio. A variedade de estruturas gramaticais, como a escrita informal e abreviações, dificultam preservar a semântica do texto (Abulaish; Sah, 2019). Sendo assim, normalmente são aplicadas tarefas de pré-processamento ao corpus textual, por exemplo, remoção de URLs e palavras acompanhadas de caracteres especiais (ex. #, @).

Além disso, cada idioma possui sua própria morfologia linguística. A morfologia do inglês é particionada em relações morfológicas flexionais, derivacionais e compostas. Palavras com mais de um sentido em inglês podem ser traduzidas em palavras totalmente diferentes em outro idioma (Miller, 1995).

Apesar dos desafios associados ao texto, existem diversas técnicas de DA para NLP, a maioria delas oferece *trade-offs*[1] entre facilidade na implementação e melhoria no desempenho do modelo (Feng *et al.*, 2021).

2.4 Técnicas de Data Augmentation para Textos

A análise das técnicas de DA demonstra que métodos não supervisionados simples e eficazes, incluindo modelos de tradução, paráfrase baseada em dicionário de sinônimos e subs-

[1] Trade-offs é uma expressão em inglês que remete a escolha de uma opção em detrimento de outra.

tituição aleatória, são bastante populares. Além disso, métodos que podem ser aprendidos como geração de modelo baseado em paráfrase e modelos pré-treinados baseados em amostragem, também recebem muita atenção por causa de sua diversidade e eficácia. De acordo com a diversidade dos dados os métodos de DA são classificados em paráfrase, ruído e amostragem (Li; Hou; Che, 2022).

Tarefas de NLP geralmente estão relacionadas à classificação de texto, geração de texto e previsão estruturada. Já os métodos de DA são comumente aplicados em tarefas de classificação de texto. A geração de texto tem preferência pelo uso de métodos baseados em amostragem, enquanto a previsão estruturada tem maior destaque para métodos de paráfrases (Li; Hou; Che, 2022).

A seguir serão apresentados na Figura 4 os métodos de DA em NLP de acordo com a taxonomia definida por (Li; Hou; Che, 2022).

2.4.1 Métodos de paráfrase

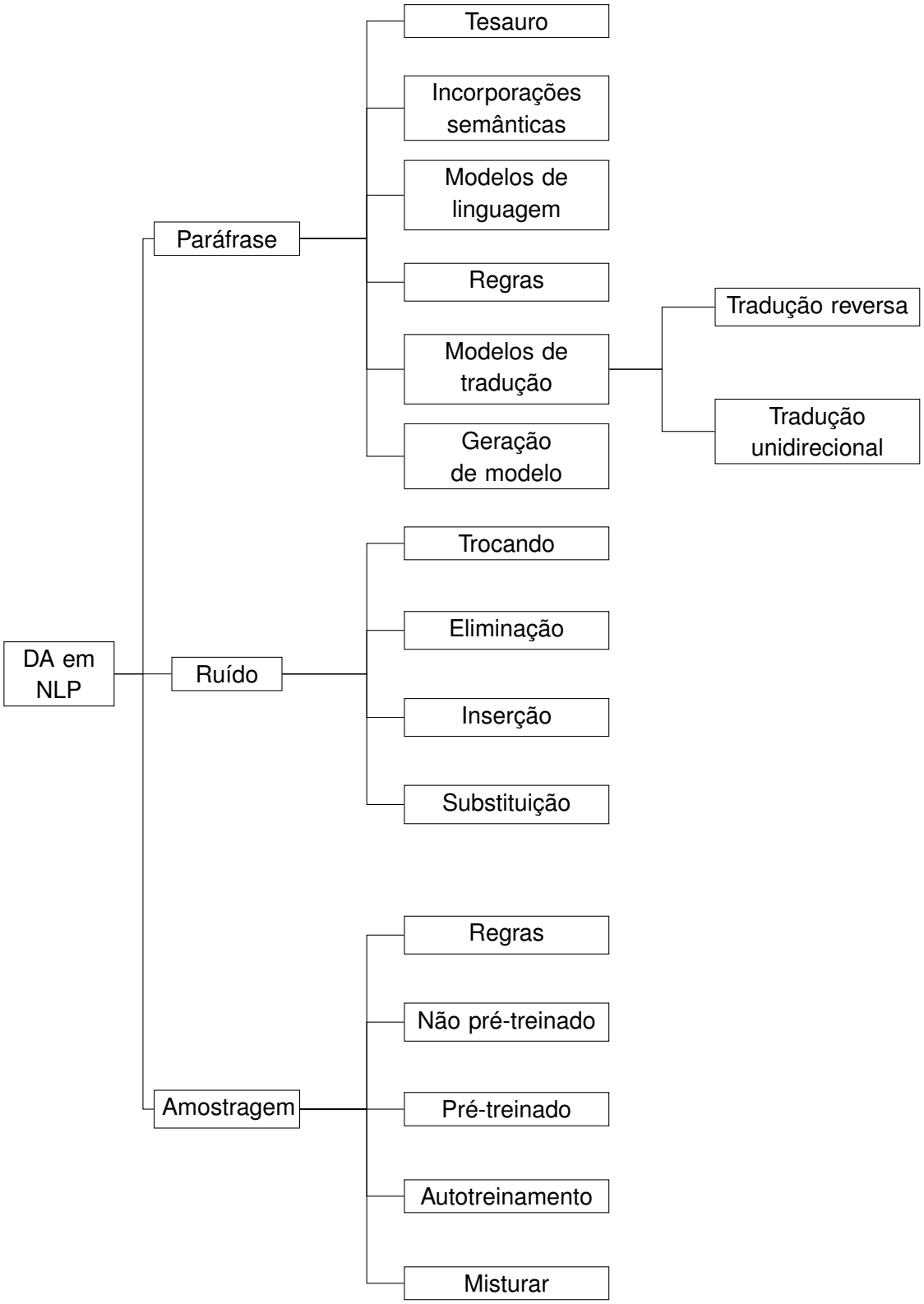
A paráfrase consiste em vários níveis, incluindo paráfrase lexical, paráfrase de frase e paráfrase de sentença. Por se adequar bem ao aumento de dados, métodos baseados em paráfrase transmitem informações muito semelhantes aos dados originais (Li; Hou; Che, 2022). No ramo da paráfrase quatro métodos de maior interesse para a proposta deste trabalho são apresentadas abaixo:

- Tesauro: em inglês *thesauruses*, ou simplesmente dicionário de sinônimos, é um método de DA que substitui palavras ou frases por seus sinônimos. A Figura 5 mostra um exemplo desse método. Um estudo proposto por (Zhang; Zhao; LeCun, 2015) obteve o tesauro do banco de dados léxico em inglês WordNet. Nesse contexto, cada sinônimo de uma palavra ou frase é classificado pela proximidade semântica ao significado mais frequente. Para decidir quantas palavras substituir, são extraidas todas as palavras substituíveis do texto fornecido e aleatoriamente algumas delas são escolhidas para serem substituídas.

- Incorporações semânticas: esse método usa incorporações de palavras pré-treinadas, substituindo no espaço de incorporação a palavra original na sentença por seu vizinho mais próximo, conforme exemplificado na Figura 6. Proposto por (Wang; Yang, 2015), um estudo utilizou esse método na tarefa de classificação de mensagens do Twitter para substituir cada palavra original em um *tweet*[2] por uma de suas k-palavras vizinhas mais próximas, com base na similaridade de cosseno entre os vetores de consulta e de palavras-alvo.

[2] *Tweets* é um termo utilizado para designar as publicações feitas na rede social do Twitter.

Figura 4 – Taxonomia dos métodos de DA em NLP.



Fonte: Adaptado de (Li; Hou; Che, 2022).

Figura 5 – Exemplo de tesauro em inglês.

Fonte: Adaptado de (Li; Hou; Che, 2022).

Figura 6 – Exemplo de incorporações semânticas em inglês.

Fonte: Adaptado de (Li; Hou; Che, 2022).

- Modelos de linguagem: são amplamente utilizados para DA devido à sua alta performance. Uma abordagem proposta por (Wu *et al.*, 2019), utilizou *Bidirectional Encoder Representations from Transformers* (BERT) condicional para aumentar as sentenças sem quebrar a compatibilidade dos rótulos. Especificamente, algumas palavras de uma sentença rotulada são mascaradas aleatoriamente e então preenchidas pelo BERT condicional. A Figura 7 apresenta um exemplo de modelo de linguagem.

Figura 7 – Exemplo de modelos de linguagem em inglês.

Fonte: Adaptado de (Li; Hou; Che, 2022).

- Modelos de tradução: os modelos de tradução são comumente aplicados em DA, visto que a tradução é um meio natural de parafrasear. Os modelos de tradução podem traduzir um texto para outro idioma e retornar ou não para o idioma original, essa característica distingue os métodos de DA apresentados a seguir.

  - **Tradução unidirecional** é um método que traduz o texto original para outros idiomas. Uma abordagem com *tweets* em inglês usou este método para aumentar os dados por meio de traduções do inglês para o francês, espanhol, alemão e italiano. Os experimentos demonstraram bons resultados para pequenos corpora de *tweets* em idiomas diferentes do inglês (Barriere; Balahur, 2020). A Figura 8 ilustra esse método, no qual o Google tradutor é utilizado para traduzir uma sentença do português para o inglês.

– **Tradução reversa**, ou em inglês *back-translation*, é um método em que o texto é traduzido para outros idiomas, e depois traduzido de volta com o intuito de obter dados aumentados no idioma original. Modelos de tradução automática treinados e serviços de *Application Programming Interface* (API)[3] de tradução em nuvem, como Google e DeepL, são ferramentas comuns para *back-translation* (Li; Hou; Che, 2022). Esse método foi empregado por (Xie *et al.*, 2020) nos idiomas inglês e francês, com o propósito de parafrasear os dados de treinamento em tarefas de classificação de texto. No exemplo da Figura 9 uma sentença é traduzida do português para o inglês, e traduzida novamente do inglês para o português com o auxílio do Google tradutor.

Figura 8 – Exemplo de tradução unidirecional.

Fonte: Autoria própria (2023).

Figura 9 – Exemplo de *back-translation*.

Fonte: Autoria própria (2023).

2.4.2 Métodos de ruídos

São métodos que adicionam pequenos ruídos para proporcionar um aumento adequado dos dados sem afetar seriamente a semântica. Assim, além de expandir a quantidade de dados de treinamento, também melhora a robustez do modelo (Li; Hou; Che, 2022). Apresentado por (Wei; Zou, 2019), o EDA usa quatro métodos de DA baseados em ruído para aumentar o desempenho em tarefas de classificação textual. A seguir são abordados os métodos de ruído propostos no EDA e de relevância para esta pesquisa.

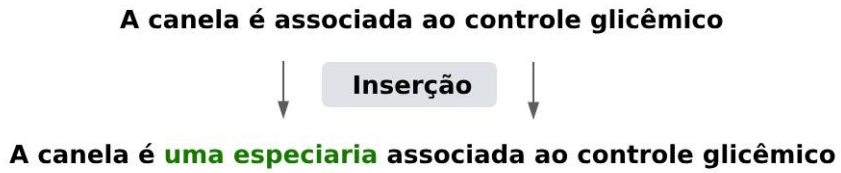
- Inserção: este método baseado em ruídos insere arbitrariamente palavras em uma frase, conforme Figura 10. Na proposta de (Wei; Zou, 2019), uma variação do método

[3] APIs são mecanismos que permitem a comunicação entre dois componentes de software usando um conjunto de definições e protocolos.

de inserção, o *random insertion* encontra um sinônimo aleatório de uma palavra e insere em uma posição aleatória na frase, esse processo é repetido n vezes.

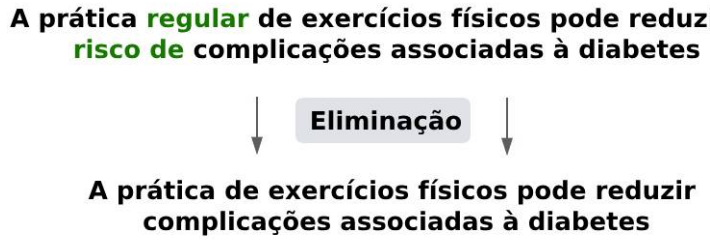
- Eliminação: o método de eliminação usa probabilidade para remover aleatoriamente palavras em uma frase (Wei; Zou, 2019). A remoção de uma palavra da sentença, sem alteração em seu significado, pode ser observada no exemplo da Figura 11.
- Trocando: do inglês *swapping*, é um método que consiste em trocar de forma aleatória a posição entre duas palavras em uma sentença e repetir o processo um determinado número de vezes (Wei; Zou, 2019). A Figura 12 apresenta um exemplo em português de duas palavras trocando suas posições na sentença.
- Substituição: é um método baseado em ruídos que escolhe de forma aleatória palavras que não são *stop word*[4] e, substitui ao acaso cada uma das palavras por seus sinônimos (Wei; Zou, 2019). Este método é demonstrado na Figura 13.

Figura 10 – Exemplo do método de inserção no idioma português.



Fonte: Autoria própria (2023).

Figura 11 – Exemplo do método de eliminação no idioma português.



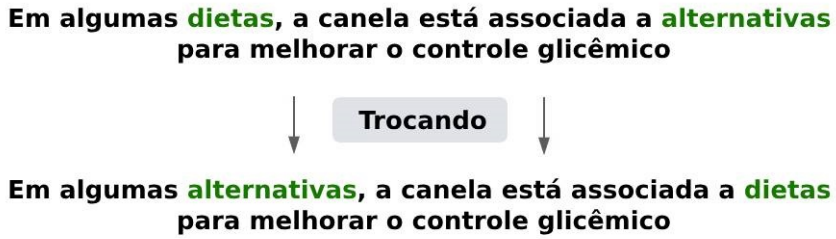
Fonte: Autoria própria (2023).

2.4.3 Métodos de amostragem

São métodos que captam uma distribuição e amostram novos dados, geralmente são específicos de tarefas, e exigem informações como rótulos e formato dos dados (Li; Hou; Che, 2022). Na sequência, serão abordados e detalhados os métodos de amostragem.

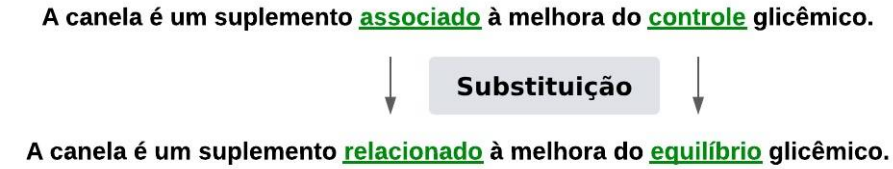
[4] *Stop words* são palavras comuns em um idioma que têm pouco significado e não interferem na semântica da sentença.

Figura 12 – Exemplo do método trocando no idioma português.



Fonte: Autoria própria (2023).

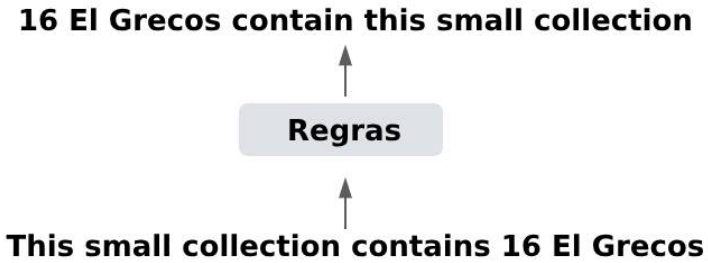
Figura 13 – Exemplo do método de substituição no idioma português.



Fonte: Autoria própria (2023).

- Regras: esses métodos têm regras específicas para gerar diretamente a expansão dos dados, como mostra a Figura 14. Por vezes os métodos de regras baseados em amostragem precisam de heurísticas de NLP e rótulos correspondentes para garantir a validação dos dados aumentados (Li; Hou; Che, 2022). No cénario das regras baseadas em amostragem, (Kang *et al.*, 2018) propõe um método guiado pelo conhecimento em regras para combinar novas sentenças com sentenças originais e gerar um aumento nos dados de treinamento do modelo.

Figura 14 – Exemplo do método de amostragem utilizando regras em inglês.

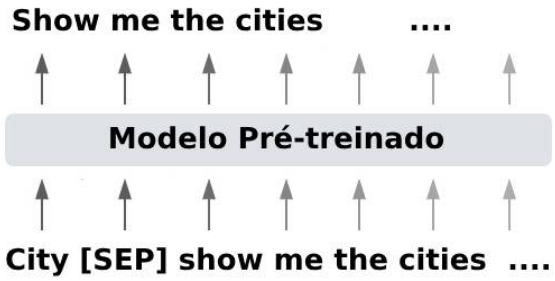


Fonte: Adaptado de (Li; Hou; Che, 2022).

- Pré-treinado: os métodos pré-treinados em DA utilizam modelos que foram ajustados em grandes conjuntos de dados de treinamento para aumentar seu próprio conjunto de dados. De acordo com (Kumar; Choudhary; Cho, 2020), durante o pré-treinamento

os modelos são treinados em *autoencoder*[5], (por ex. BERT) ou em auto-regressão[6] (por ex. GPT2). Dentro da configuração *autoencoder*, determinados *tokens* são mascarados na frase e o modelo prevê esses *tokens*. Dentro de uma configuração de auto-regressão, o modelo prevê a próxima palavra para um dado contexto. Uma representação dessa abordagem pode ser visualizada na Figura 15.

Figura 15 – Exemplo de DA com método pré-treinado em inglês.



Fonte: Adaptado de (Li; Hou; Che, 2022).

- Autotreinamento: o autotreinamento é um método de DA composto por um modelo professor treinado com dados rotulados que cria rótulos sintéticos para exemplos não rotulados, conforme Figura 16. Os dados sintéticos criados são usados para treinar um modelo aluno. Esse método é denominado autotreinamento quando o modelo aluno adquire capacidade semelhante ou superior ao modelo professor (Du *et al.*, 2020). (Miao; Last; Litvak, 2020) usou o método de autotreinamento em uma abordagem com destilação de dados no processo de autotreinamento, objetivando aumentar os dados do Twitter e monitorar a opnião pública sobre as medidas de intervenção tomadas pelo governo durante o COVID-19.

Figura 16 – Exemplo com autotreinamento em inglês.

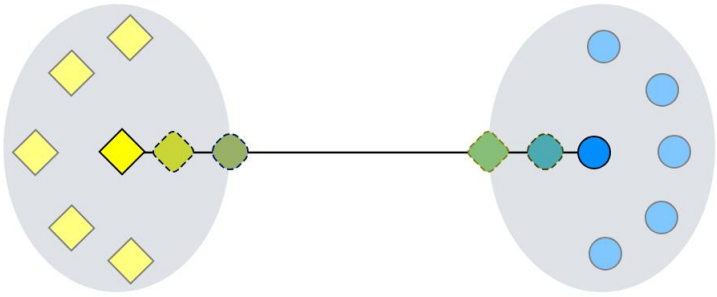
Fonte: Adaptado de (Li; Hou; Che, 2022).

[5] Autoencoder é um meio para aprender codificações eficientes de dados não rotulados.

[6] Auto-regressão prevê resultados futuros de uma sequência a partir da observação de uma sequência anterior.

- Misturar: do inglês *mixup*, é um método que gera novas amostras por interpolação linear de várias amostras e seus rótulos. (Sun *et al.*, 2020) introduz uma abordagem que combina o método *mixup* com um transformador baseado em arquitetura pré-treinada para testar seu desempenho em conjuntos de dados de classificação textual. O exemplo observado na Figura 17 mostra dois conjuntos, um representando por amostras em amarelo e outro em azul. O processo do *mixup* seria escolher aleatoriamente dois exemplos, um de cada conjunto, e criar uma nova amostra combinando as características desses dois exemplos. Essa combinação é realizada pela interpolação linear das amostras.

Figura 17 – Exemplo do método misturar.



Fonte: Adaptado de (Li; Hou; Che, 2022).

3 TRABALHOS RELACIONADOS

Na perspectiva dos métodos de DA em NLP, diversas abordagens colaboram para o desenvolvimento de pesquisas futuras. Dessa forma, este capítulo apresenta os trabalhos relacionados que posicionaram o presente estudo no contexto amplo da literatura sobre métodos de DA.

3.1 Método de incorporações semânticas aplicado ao Twitter

A abordagem proposta por (Wang; Yang, 2015) usa incorporações semânticas como método de DA para aprimorar a análise comportamental computacional usando texto de mídia social. O estudo extraiu um corpus contendo 3.375 *tweets* com comportamentos irritantes associados a *hashtag* #petpeeve. As categorias dos comportamentos nos *tweets* são rotuladas usando o modelo *Latent Dirichlet Allocation* (LDA) *clustering*[1] associado à *human-identification*.

A análise dos dados usa *Sparse Additive Generative Model* (SAGE) para visualizar palavras salientes em cada categoria dos comportamentos irritantes nos *tweets* com *hashtag* #petpeeve (Wang; Yang, 2015).

O método *GoogleNews* de incorporações lexicais treinado com 100 bilhões de palavras foi uma das abordagens que (Wang; Yang, 2015) usou nos experimentos. Esse método demonstrou uma melhoria de 6.1% em relação à linha de base sem aumento de dados.

De acordo com o autor, o uso de incorporações lexicais e semânticas como método de DA melhoram significativamente o desempenho para categorização automática de comportamentos irritantes (Wang; Yang, 2015).

3.2 Aumento contextual com BERT condicional

Proposto por (Wu *et al.*, 2019), o BERT condicional realiza o aumento contextual das sentenças sem quebrar a compatibilidade dos rótulos. O método aplica o aumento contextual por BERT condicional, e é ajustado com BERT. De acordo com o autor, o BERT foi adotado como modelo de pré-treino porque é um método bidirecional poderoso, e por ser baseado em tranformador torna a memória mais estruturada para lidar com dependências de longo prazo no texto.

(Wu *et al.*, 2019) apresentou o *Conditional Masked Language Model* (C-MLM), este modelo condicional mascara aleatoriamente alguns dos *tokens* de uma entrada, objetivando prever uma palavra rotulada compatível com base no contexto e no rótulo da frase. O C-MLM oferece uma solução para modelos de linguagem mascarada, pois além do contexto, também considera a compatibilidade da palavra prevista com os rótulos anotados das sentenças originais.

[1] *Clustering* é uma coleção de objetos que são “semelhantes” entre eles e são “dissimilares” aos objetos pertencentes a outros *clusters*(Madhulatha, 2012).

Os resultados mostram que, para vários conjuntos de dados em diferentes arquiteturas de classificadores, o aumento contextual do BERT condicional apresenta melhores resultados no desempenho do modelo. Segundo (Wu *et al.*, 2019), os experimentos foram realizados em seis diferentes tarefas de classificação de texto e, demonstrou que o modelo do BERT condicional é superior no aumento de sentenças má comparação com as linhas de base. No estudo os métodos com sinônimos (Miller, 1995), com contexto (Kobayashi, 2018) e, com contexto e rótulos (Kobayashi, 2018) foram comparados ao BERT condicional. Além disso foram usados seis datasets de classificação, conforme Tabela 1.

3.3 QANet com método *back-translation*

Uma nova arquitetura para o modelo *Question Answering* (Q&A) chamada QANet, usa convolução para modelar interações locais e auto-atenção em interações globais. No conjunto de dados SQUAD, o QANeT é mais rápido em treinamento e em inferência, além de alcançar precisão equivalente a modelos recorrentes. A rapidez no treinamento do modelo permite aumentar o conjunto de dados. Dessa forma, o método de *back-translation* é incorporado ao modelo para aumentar a diversidade nos dados de treinamento (Yu *et al.*, 2018).

Em relação ao dataset, o *Stanford Question Answering Dataset* (SQuAD) contém 107.700 pares de Q&A, com 87.500 para treinamento, 10.100 para validação e outros 10.100 para teste. Visando generalizar a eficiência do modelo para outros datasets, os experimentos também foram replicados no conjunto de dados TriviaQA. Para o pré-processamento dos dados o estudo usou a biblioteca *Natural Language Toolkit* (NLTK) (Yu *et al.*, 2018).

De acordo com (Yu *et al.*, 2018), o método de DA proposto pode trazer melhorias não triviais em termos de precisão além de ser aplicável em outras tarefas de NLP, especialmente quando os dados de treinamento são insuficientes. Os resultados do experimento demonstraram que o modelo treinado com DA obteve um ganho significativo de 1.5/1.1 nas métricas de avaliação de precisão EM/F1.

3.4 Análise dos trabalhos relacionados

Assim como a proposta de (Wang; Yang, 2015), o presente trabalho obteve o dataset do Twitter, porém utilizando-se de parâmetros voltados a diabetes. *Tweets* geralmente contemplam textos curtos e com muito ruído, sendo necessário realizar o pré-processamento dos dados.

Conforme o trabalho proposto por (Yu *et al.*, 2018), o método *back-translation* em DA apresentou bons resultados. De acordo com (Federation, 2021) a diabetes afeta pessoas no mundo inteiro, desta forma, é importante considerar abordagens de DA que aumentem o conjunto de dados em outros idiomas. Dessa forma, o método *back-translation* pode ser aplicado independente de linguagem.

Além dos trabalhos relacionados apresentados neste capítulo, outros métodos de DA propostos possuem abordagens interessantes para o aumento do corpus sobre diabetes. (Coulombe, 2018) apresentou um método de tesauro que além de sinônimos, usa hiperônimos[2] para substituir as palavras originais. A partir de dicionários de pares de palavras, (Regina; Meyer; Goutal, 2020) realizou substituições entre a forma expandida e abreviada das palavras. O método proposto por (Hou *et al.*, 2018) aproveitou as mesmas alternativas semânticas de um enunciado em dados de treinamento com modelos *Sequence-to-Sequence* (Seq2Seq)[3] para compreensão da linguagem em um sistema de diálogo orientado a tarefas. Os trabalhos relacionados podem ser observados a seguir no Quadro 1.

Quadro 1 – Estudos envolvendo métodos de DA em NLP

Método de DA	**Título**	**Dataset**
Thesauruses (Coulombe, 2018)	Text Data Augmentation Made Simple By Leveraging NLP Cloud APIs	Polarity v2.0
Semantic Embeddings (Wang; Yang, 2015)	That's So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using #petpeeve Tweets	Twitter corpus
Language Models (Wu *et al.*, 2019)	Conditional BERT Contextual Augmentation	SST-5; SST-2; SUBJ; MPQA; RT; TREC
Rules (Regina; Meyer; Goutal, 2020)	Text Data Augmentation: Towards better detection of spear-phishing emails	SST-2; TREC-6; BEC
Back-translation (Yu *et al.*, 2018)	QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension	SQuAD; TriviaQA
Model Generation (Hou *et al.*, 2018)	Sequence-to-Sequence Data Augmentation for Dialogue Language Understanding	ATIS
Noising-based (Wei; Zou, 2019)	EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks	SST-2; CR; SUBJ; TREC; Pro-Con
Pretrained Models (Zhang *et al.*, 2020)	On Data Augmentation for Extreme Multi-label Classification	AmazonCat-13K

Fonte: Autoria própria (2023).

[2] Hiperônimo é toda palavra que possui sentido amplo.

[3] Seq2Seq é uma abordagem de ML que recebe uma sequência como entrada, em um domínio específico, e converte essa sequência para uma representação em outro domínio.

4 MATERIAIS E MÉTODOS

4.1 Materiais

Os materiais utilizados foram provenientes da revisão da literatura, de estudos similares previamente realizados e de tendências em recursos computacionais. Esses recursos computacionais são extremamente importantes, pois proporcionam um meio de aplicar os conhecimentos teóricos adquiridos na pesquisa.

As principais fontes de pesquisa empregadas para explorar a vasta gama de artigos científicos que contribuiram no desenvolvimento desse trabalho foram as plataformas IEEE, Google Scholar e arXiv.

Os códigos foram implementados na linguagem de programação Python - versão 3.6.9, disponível por meio do Google Colaboratory. Colaboratory ou “Colab”, como é mais conhecido. Trata-se de um produto do Google Research, área de pesquisas científicas da Google. O Colab permite que qualquer pessoa escreva e execute código Python arbitrário pelo navegador e é especialmente adequado para ML, análise de dados e educação. Mais tecnicamente, o Colab é um serviço de notebooks hospedados do Jupyter que não requer nenhuma configuração para usar e oferece acesso sem custo financeiro a recursos de computação como *Graphics Processing Units* (GPUs).

Python é uma linguagem de programação de alto nível com propósito geral, amplamente empregada em ML. Dotada de bibliotecas multifuncionais, destaca-se no processamento de textos. Além disso, a integração eficiente com APIs de diversas plataformas é uma característica proeminente dessa linguagem. Nesse contexto, o presente trabalho adotará as seguintes bibliotecas, APIs, modelo pré-treinado e métodos:

- **NLTK** : conjunto de bibliotecas e programas para processamento simbólico e estatístico da linguagem natural;
- **Sentence-transformers**: biblioteca de ferramentas para criar representações numéricas de sentenças e documentos de texto, conhecidas como vetores de *embeddings*;
- ***Robustly Optimized BERT Approach*** **(RoBERTa)**: é um modelo pré-treinado que tem a proposta de aprimorar o procedimento de pré-treinamento do BERT. Esse aperfeiçoamento ocorre com modificações simples que incluem: treinar o modelo por mais tempo, com lotes maiores e em mais dados; remover o objetivo de previsão da próxima frase; treinar em sequências mais longas; e mudar dinamicamente o padrão de mascaramento aplicado aos dados de treinamento (Liu *et al.*, 2019). O *stsb-distilroberta-base-v2* é uma versão compacta do RoBERTa, esse modelo de transformadores permite mapear sentenças para um espaço vetorial denso de 768 dimensões, e pode ser usado para tarefas como agrupamento ou pesquisa semântica (Reimers; Gurevych, 2019);

- **Bokeh**: é uma biblioteca de visualização em Python usada para criar visualizações de dados e painéis de controle interativos;
- ***t-Distributed Stochastic Neighbor Embedding* (t-SNE)**: é um método utilizado para visualizar dados de alta dimensão, atribuindo a cada ponto um posicionamento em um mapa bidimensional ou tridimensional (Maaten; Hinton, 2008). A visualização dos dados em um espaço de menor dimensão pode facilitar a identificação de *clusters*, padrões e estruturas nos dados. O trabalho proposto em Risk Factor Mining: CORD-19 Articles 2021, foi utilizado como referência para o uso do t-SNE pois a abordagem apresenta uma interface interativa que permite visualizar cada ponto do gráfico de forma independente;
- **Pandas**: biblioteca de software criada para a linguagem Python, a fim de fazer manipulação e análise de dados;
- **Matplotlib**: é uma biblioteca amplamente utilizada em Python para criação de gráficos e visualizações de dados;
- **Textaugment**: biblioteca de ferramentas usada para aumentar ou gerar variações de texto de entrada (Marivate; Sefara, 2020);
- **Gensim**: é uma biblioteca de NLP em Python que se concentra em modelagem de tópicos e incorporação de palavras;
- **Tensorflow**: biblioteca originalmente desenvolvida para Python voltada para aplicações de ML e DL;
- **Scikit-learn**: biblioteca de ML para a linguagem de programação Python.
- **K-means**: é um método de *clustering*[1] disponível na biblioteca scikit-learn. Cada um dos *clusters* é representado pela média (geralmente ponderada) de seus pontos, o chamado centróide. Além disso, requer que o número *k* de *clusters* seja especificado previamente (Rai; Shubha, 2010);
- **K-Fold**: é uma técnica de validação cruzada disponível como uma função da biblioteca *scikit-learn*. A técnica consiste em dividir todas as amostras em grupos de amostras, chamados *folds*, de tamanhos iguais. Um modelo é treinado usando **k-1** dos *folds* como dados de treinamento, o modelo resultante é validado com a parte restante dos dados, denominada conjunto de testes, a fim de calcular uma medida de desempenho, como por exemplo, a *accuracy* (Pedregosa *et al.*, 2011);
- **Cleantext**: biblioteca para realizar a limpeza de dados brutos de texto em Python.

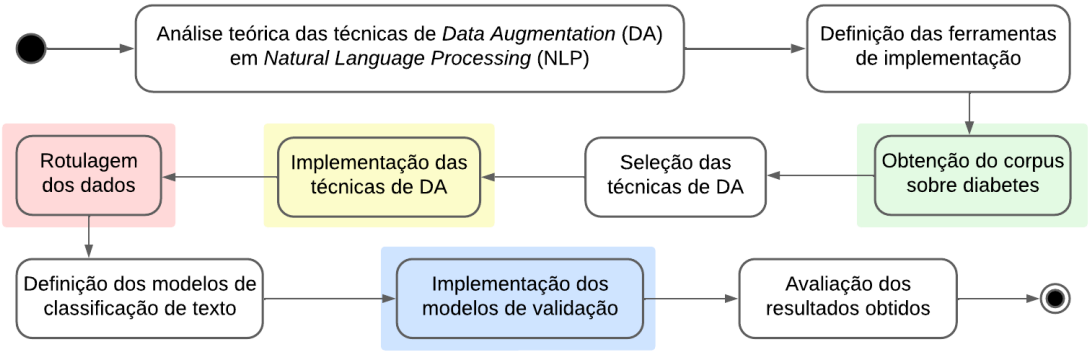
[1] *Clustering* é a divisão de dados em grupos de dados semelhantes. Cada grupo, chamado *cluster*, consiste em objetos que são semelhantes entre si e diferentes de objetos de outros grupos.

- **Googletrans**: API de tradução do Google Translate em Python;
- **Tweepy**: é uma biblioteca Python que simplifica a interação com a API do Twitter.

4.2 Métodos

Este trabalho aborda a validação de técnicas de DA em corpus textuais sobre diabetes com o objetivo de promover uma base de qualidade para tarefas em NLP na área da saúde. Dessa forma, o trabalho proposto segue as etapas dispostas no fluxograma metodológico geral representado na Figura 18.

Figura 18 – Fluxograma metodológico geral.



Fonte: Autoria própria (2023).

1. **Análise teórica das técnicas de DA em NLP**: essa etapa compreende o embasamento teórico sobre as técnicas de DA existentes na literatura para abordagens em NLP. Isso foi necessário para analisar as melhores estratégias de implementação, vantagens e desvantagens do uso de cada técnica;

2. **Definição das ferramentas de implementação**: etapa responsável por definir os materiais necessários para o desenvolvimento do trabalho. Nesse momento, foram definidos: a linguagem de programação, bibliotecas, APIs e plataformas usadas no processo de implementação. Essas decisões foram orientadas por necessidades específicas, considerando como principais fatores a disponibilidade dos recursos e a capacidade de atender aos objetivos do projeto;

3. **Obtenção do corpus sobre diabetes**: o foco dessa etapa é criar um corpus textual público abordando o tema da diabetes. Para tal, seguiu-se com uma abordagem similar à empregada por (Karami *et al.*, 2018), representada na Figura 19. A API *Tweepy* versão 4.12.1 disponibiliza parâmetros que permitem refinar a busca de *tweets* com base em palavras-chave. Esses parâmetros são escolhidos de acordo com as características desejadas nos *tweets*, possibilitando escolher o idioma, a string de consulta, a

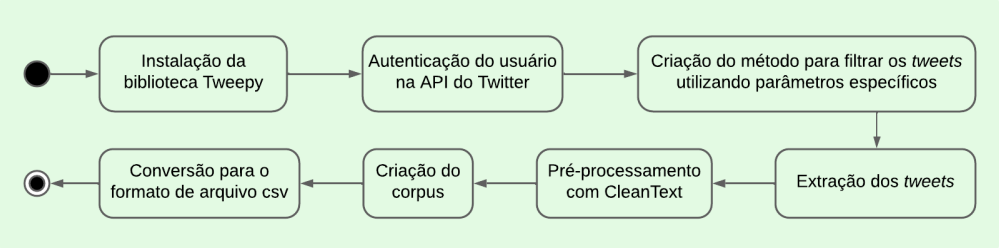
localização do usuário, dentre outros parâmetros que são informados na documentação. Foram extraídos apenas *tweets* relacionados a diabetes, todos no idioma inglês. O inglês é um dos idiomas predominantes no mundo, portanto espera-se um maior número de *tweets* nesse idioma.

O estudo utilizou o nível de acesso gratuito da API do Twitter para realizar a extração dos *tweets*. Em abril de 2023, o Twitter anunciou a alteração dos valores para cada nível de acesso. Atualmente, os níveis de acesso são divididos em quatro: gratuito, básico, pro e empresarial. O custo de cada nível de acesso está relacionado diretamente à quantidade de recursos disponíveis na API.

Geralmente textos derivados de mídias sociais apresentam diversos ruídos, como símbolos e caracteres especiais, hiperlinks, emojis, espaços vazios, dentre outros elementos que dificultam manter a semântica no texto. Portanto, foi realizado um pré-processamento dos *tweets* com a versão 0.6.0 da biblioteca *Cleantext*, objetivando a remoção desses ruídos.

Após a extração e pré-processamento dos *tweets*, o corpus foi convertido em um formato de arquivo *Comma-Separated Values* (CSV)[2] e salvo em um diretório no Google drive.

Figura 19 – Fluxograma da obtenção do corpus.



Fonte: Autoria própria (2023).

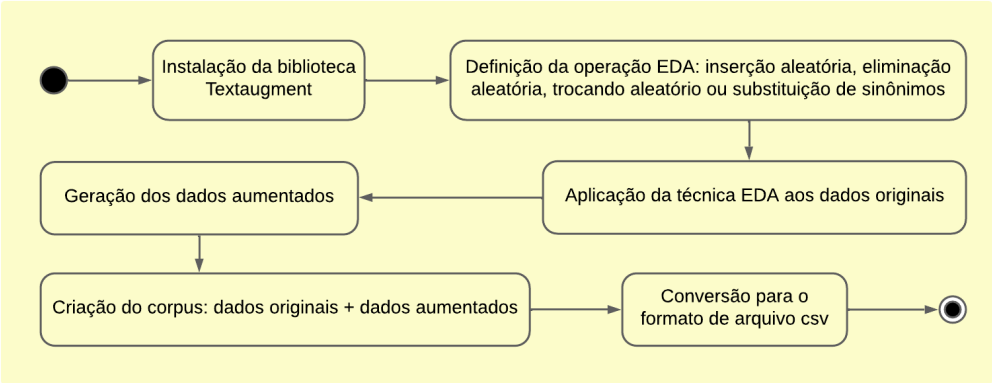
4. **Seleção das técnicas de DA**: de acordo com as características mais pertinentes à proposta, duas técnicas de DA foram selecionadas, sendo estas EDA e *back-translation*. A técnica EDA foi selecionada por apresentar bom desempenho e conservar o significado das sentenças aumentadas em tarefas de classificação de texto, conforme demonstrado por (Wei; Zou, 2019). Por sua vez, a técnica de *back-translation* demonstra relevância em diversos estudos, como (Luque, 2019), (Ibrahim; Torki; El-Makky, 2020), (Aroyehun; Gelbukh, 2018) que utilizaram a técnica em corpus obtidos de mídias sociais, similares ao Twitter;

5. **Implementação das técnicas de DA**: etapa de implementação das técnicas de DA aplicadas ao corpus sobre diabetes. Técnicas como EDA e *back-translation* possuem

[2] CSV é um arquivo de texto com formato específico para possibilitar o salvamento dos dados em um formato estruturado de tabela.

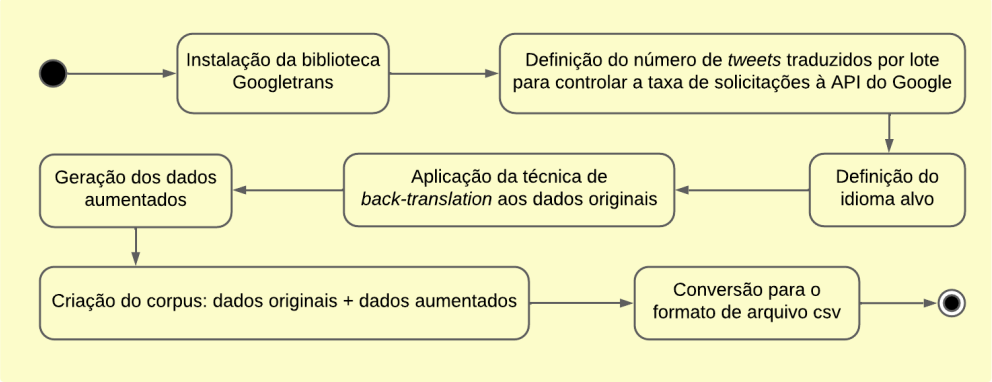
o potencial de enriquecer o corpus, tornando-o mais robusto e adequado para análises detalhadas. A representação visual da implementação das técnicas EDA e *back-translation* pode ser observada nas Figuras 20 e 21, respectivamente;

Figura 20 – Fluxograma de implementação da técnica EDA



Fonte: Autoria própria (2023).

Figura 21 – Fluxograma de implementação da técnica *backtranslation*



Fonte: Autoria própria (2023).

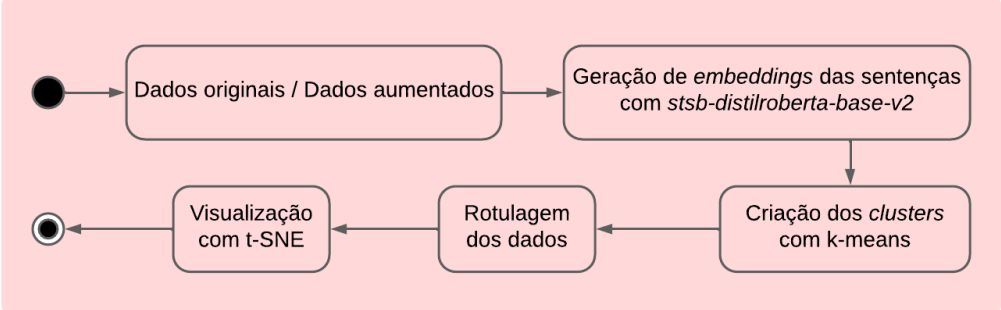
6. **Rotulagem dos dados**: a rotulagem dos dados se fez necessária para que fosse possível aplicar a base de dados original e original + aumentada aos métodos de classificação, a fim de identificar eventuais melhorias nas métricas de avaliação, resultantes da implementação das técnicas de DA. Os *tweets* são rotulados, conforme visualizados na Figura 22.

   Capturar informações semânticas e estruturais de dados textuais requer o uso de *embeddings*, que podem ser obtidos de palavras ou de sentenças, e que definem suas representações vetoriais. A geração dos *embeddings* possibilita a aplicação do método de *clustering k-means*[3]. O objetivo da aplicação do *k-means* é agrupar *tweets*

[3] O *k-means* é um algoritmo de agrupamento de dados que tenta separar amostras em n grupos de variância igual, minimizando um critério conhecido como inércia que diz quão coerentes internamente são os *clusters* (Pedregosa *et al.*, 2011).

similares com relação ao seu contexto, e assim verificar se os textos agrupados fazem sentido juntos. Cada *cluster* é então definido como uma categoria, enumerada de 0 a 5. Por fim, para facilitar a visualização dos *clusters* formados, é empregado o método t-SNE;

Figura 22 – Fluxograma do processo de rotulagem dos dados.



Fonte: Autoria própria (2023).

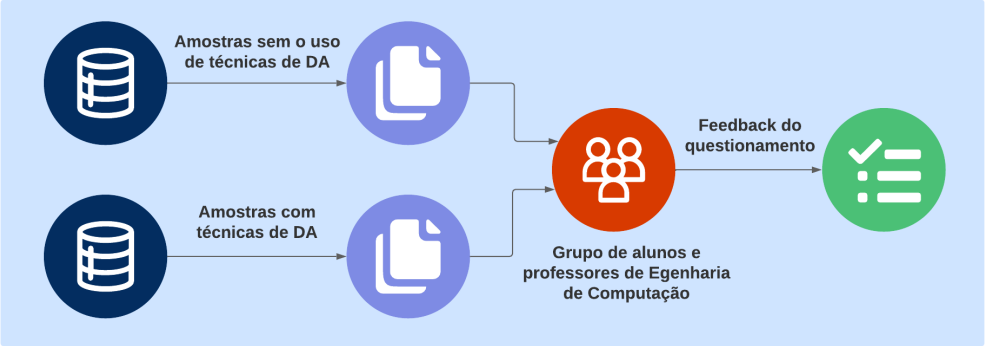
7. **Definição dos modelos de classificação de texto**: a classificação de texto consiste na extração de características de dados de textos brutos e, a partir dessas características, realizar a previsão das categorias dos dados textuais (Li *et al.*, 2021). Para realizar uma comparação de desempenho na classificação de texto em cada técnica de DA proposta, foram selecionados os seguintes modelos:

    - ***Naïve Bayes* (NB)**: é um classificador probabilístico relacionado ao teorema de Bayes, com uma forte suposição de independência entre as características. A vantagem do classificador NB é o menor tempo computacional necessário para treinar os dados (N; Gupta, 2020b).
    - ***Support Vector Machine* (SVM)**: este modelo constrói um hiperplano no espaço de entrada unidimensional ou espaço de características, maximizando a distância entre o hiperplano e as duas categorias de conjuntos de treinamento, alcançando uma melhor capacidade de generalização (Li *et al.*, 2021).
    - ***Random Forest* (RF)**: este modelo é uma combinação de árvores de decisão, tais que, cada árvore depende dos valores de um vetor aleatório amostrado de forma independente e com mesma distribuição para todas as árvores da floresta (Breiman, 2001).

8. **Implementação dos modelos de validação**: essa etapa adota abordagens para avaliar a eficácia das técnicas de DA. Esse processo de validação é conduzido empregando duas abordagens distintas:

- **Validação por meio de métricas de desempenho quantitativas**: essa abordagem emprega uma validação cruzada k-fold, onde a métrica de *accuracy*[4] foi calculada em cada iteração. Nesse contexto, o valor médio da *accuracy* possibilita avaliar o desempenho do modelo de classificação fazendo uso dos dados aumentados e dos dados originais, o que permite saber se as técnicas de DA foram efetivas.
- **Validação manual por questionamento direto**: abordagem em que os textos gerados pelas técnicas de DA foram submetidos a um processo de validação manual, por meio do questionamento direto a alunos do curso de Engenharia de Computação. Os participantes foram selecionados, a fim de saber se uma sentença original apresenta o mesmo significado que uma sentença gerada pelo DA. Uma ilustração dessa abordagem pode ser visualizada na Figura 23.

Figura 23 – Validação por questionamento direto



Fonte: Autoria própria (2023).

9. **Análise comparativa dos resultados obtidos:** etapa final do trabalho, em que os resultados obtidos com os experimentos foram comparados e analisados. Nessa etapa, métricas de desempenho aplicadas a tarefas de classificação, foram analisadas, para definir a eficiência dos métodos de EDA e *back-translation* nos corpora (original e aumentado).

[4] A *accuracy* descreve até que ponto os rótulos previstos estão de acordo com o rótulos verdadeiros. Os rótulos previstos correspondem aos rótulos das classes onde as novas instâncias são agrupadas (Ahmed; Seraj; Islam, 2020).

5 RESULTADOS

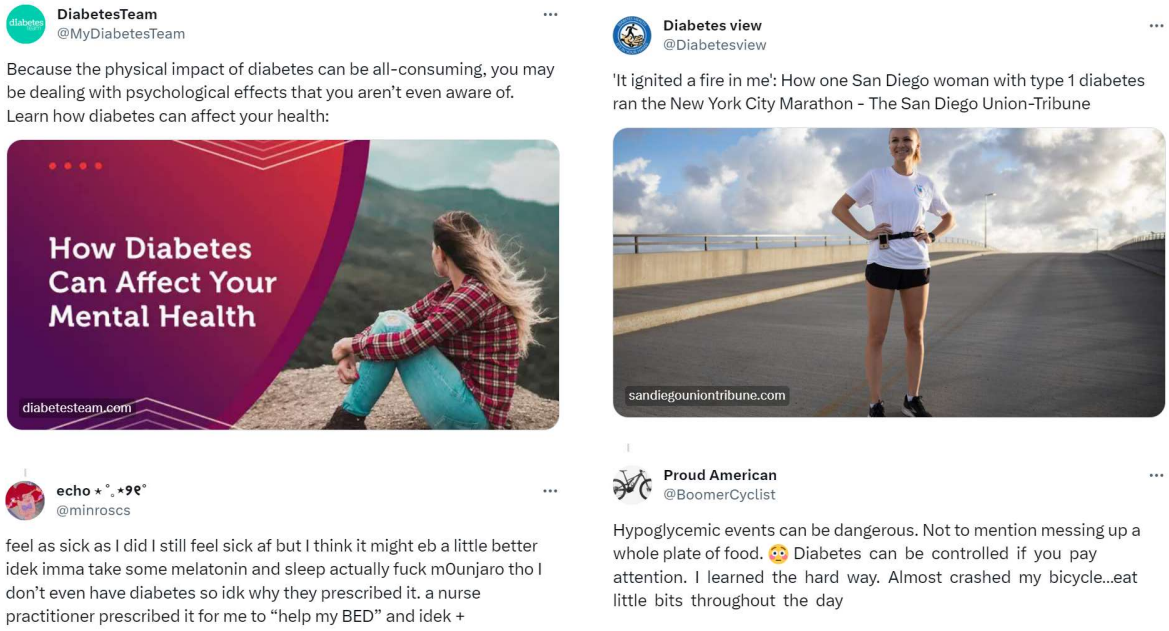
5.1 Dataset

As mídias sociais fornecem uma plataforma para os usuários expressarem suas opiniões e compartilharem informações. Opiniões de saúde nas mídias sociais, como o Twitter, oferecem uma abordagem para caracterizar questões de saúde como o diabetes (Karami *et al.*, 2018).

O dataset experimental foi criado com 2.000 *tweets* em inglês contendo as palavras "diabetes"e "#diabetes". A extração dos *tweets* foi feita com a API *tweepy* versão 4.12.1. Além disso, no processo de extração os *retweets* foram removidos para evitar dados repetidos. A API *tweepy* disponibiliza parâmetros que permitem refinar a busca dos *tweets*, como por exemplo, escolher o idioma e buscar palavras ou expressões específicas. A Figura 24 mostra alguns *tweets* pertencentes ao dataset original, retirados da página oficial do Twitter.

Normalmente os *tweets* são textos curtos e com muito ruído, sejam emojis, links URL, dentre outros caracteres que não contribuem para a semântica do texto. Esses ruídos podem ser observados em destaque no Quadro 2, onde ainda não ocorreu o pré-processamento dos *tweets*.

Figura 24 – Tweets pertencentes ao dataset.



Fonte: Autoria própria (2023).

Portanto, para remover parte do ruído e estruturar melhor os dados foi utilizada a biblioteca *Cleantext*. Parte do dataset pré-processado pode ser visualizado no Quadro 3.

Os marcadores amarelos evidenciados no Quadro 3 indicam os ruídos presentes nos *tweets* antes do pré-processamento.

Quadro 2 – Amostra do dataset com *tweets* sem pré-processamento

Tweet id	Text
1599184583680851968	feel as sick as I did I still feel sick af but I think it might eb a little better idek imma take some melatonin and sleep actually fuck m0unjaro tho I don't even have diabetes so idk why they prescribed it. a nurse practitioner prescribed it for me to "help my BED" and idek +
1599184345969102848	' It ignited a fire in me': How one San Diego woman with type 1 diabetes ran the New York City Marathon - The San Diego Union-Tribune https://t.co/b5EtHNHYEP
1599184234849370112	Hypoglycemic events can be dangerous. Not to mention messing up a whole plate of food. 😳 Diabetes can be controlled if you pay attention. I learned the hard way. Almost crashed my bicycle. . . eat little bits throughout the day
1599184207053574145	Because the physical impact of diabetes can be all-consuming, you may be dealing with psychological effects that you aren't even aware of. Learn how diabetes can affect your health: https://t.co/SD9GGQYl3F

Fonte: Autoria própria (2023).

Quadro 3 – Amostra do dataset com pré-processamento dos *tweets*

Tweet id	Text
1599184583680851968	feel as sick as I did I still feel sick af but I think it might eb a little better idek imma take some melatonin and sleep actually fuck m0unjaro tho I don't even have diabetes so idk why they prescribed it. a nurse practitioner prescribed it for me to "help my BED"and idek +
1599184345969102848	It ignited a fire in me': How one San Diego woman with type 1 diabetes ran the New York City Marathon - The San Diego Union-Tribune
1599184234849370112	Hypoglycemic events can be dangerous. Not to mention messing up a whole plate of food. Diabetes can be controlled if you pay attention. I learned the hard way. Almost crashed my bicycle...eat little bits throughout the day
1599184207053574145	Because the physical impact of diabetes can be all-consuming, you may be dealing with psychological effects that you aren't even aware of. Learn how diabetes can affect your health:

Fonte: Autoria própria (2023).

Após a extração, é realizada a criação de um dataframe contendo os *tweets*. Este dataframe é então convertido em um arquivo no formato CSV e salvo em nuvem.

5.2 Técnicas de DA

Nesta seção, são apresentados os resultados com as amostras geradas a partir das técnicas EDA e *back-translation*.

Originalmente o dataset era composto por um total de 2.000 *tweets* coletados do Twitter. No entanto, após a aplicação das técnicas de DA esse número foi ampliado para 4.000 *tweets*.

Nos Quadros 4 e 5, as marcações amarelas evidenciam as modificações nos *tweets* após a aplicação das técnicas de DA.

5.2.1 EDA

Os métodos baseados em ruído, como o EDA, tem a vantagem de melhorar a robustez do modelo. Uma das desvantagens é a interpretação inadequada do contexto, pois ao adicionar ruídos de forma aleatória a capacidade de interpretação da sentença é comprometida. De acordo com (Li; Hou; Che, 2022), os humanos conseguem adicionar ruídos com maior facilidade devido a sua compreensão semântica da linguística e seus conhecimentos prévios que permitem adicionar o ruído na sentença preservando o contexto, o que não ocorre de forma aleatória. Outra desvantagem está relacionada à limitação da diversidade, pois a adição de ruídos pode não ser suficiente para ampliar a diversidade linguística da sentença.

O experimento EDA utilizou a biblioteca em Python *textaugment*. Essa biblioteca não possui parâmetros para configurar a quantidade de ruídos no texto com as funções de EDA, pois ocorrem de forma aleatória. Amostras de *tweets* das quatro operações básicas da técnica EDA foram escolhidas aleatoriamente para demonstrar o resultado obtido em cada uma. Desta forma, foi possível analisar se o contexto das sentenças permanece o mesmo após o uso das técnicas de EDA.

O Quadro 4 mostra cinco *tweets* sem DA, em seguida os mesmos *tweets* são exibidos com as técnicas de EDA de inserção aleatória, eliminação aleatória, trocando aleatório e substituição de sinônimos, respectivamente. Ao analisar o resultado, foi observado que a aleatoriedade dessas técnicas causa uma perda de contexto das sentenças, com exceção da técnica de substituição de sinônimos que apresentou os melhores resultados. Isso ocorre porque, ao substituir uma palavra por seu sinônimo, espera-se que o contexto permaneça o mesmo. Por outro lado, em outras operações, como a eliminação aleatória, a remoção de uma palavra importante na sentença pode resultar na alteração do contexto.

5.2.2 Back-translation

Conforme observado na taxonomia da Figura 4, *back-translation* é uma técnica de DA que faz parte das máquinas de tradução. Elas apresentam uma série de vantagens, incluindo sua ampla gama de aplicações, facilidade de uso, garantia de uma sintaxe correta e uma semântica inalterada. Entretanto, apresenta limitações relacionadas à falta de flexibilidade no controle das traduções e diversidade limitada devido aos modelos fixos de tradução automática (Li; Hou; Che, 2022).

O experimento com *back-translation* utilizou a API de tradução do Google, disponível na biblioteca Python *googletrans*. Apesar da qualidade das traduções, esta API apresenta uma limitação em relação ao tempo de execução para grandes volumes de dados. A alternativa encontrada para contornar esse problema foi definir uma quantidade de 50 *tweets* traduzidos a cada iteração e pausar por 1 segundo antes de prosseguir com a técnica. Dessa forma, é

Quadro 4 – Amostra de *tweets* para demonstração da técnica EDA

Sem o uso de técnicas de DA
i shared something on fb about covid being vascular and it was news to a friend, who has diabetes, and has had covid.
research suggests vitamin d reduces blood sugar in folks with diabetes
corporate greed, lack of ethics sent us on the trajectory of generational diabetes.
if you're dealing with diabetes burnout, you're not alone. read these stories!
body mass index versus surrogate measures of central adiposity as independent predictors of mortality in type 2 diabetes

Técnica EDA com inserção aleatória
i shared something on fb about covid being shared out vascular and it was news to a friend, who has diabetes, and has had covid.
research suggests vitamin scratch d reduces blood sugar in folks with diabetes
corporate greed, lack of ethics sent flight us on the trajectory of generational diabetes.
if you're dealing with diabetes take burnout, you're not alone. read these stories!
body mass index versus surrogate power measures of central adiposity as independent predictors of mortality in type 2 diabetes

Técnica EDA com eliminação aleatória
i shared something on fb about covid being vascular and was news to a friend, who has diabetes, has had covid.
suggests vitamin d reduces blood sugar in folks with diabetes
corporate lack ethics us on the trajectory of generational diabetes.
if you're dealing with diabetes you're not alone. read these stories!
body mass index surrogate measures of central adiposity as independent predictors of mortality in type 2 diabetes

Técnica EDA trocando aleatório
i shared something on fb about covid being had and it was news to a friend, who has diabetes, and has vascular covid.
research suggests sugar d reduces blood vitamin in folks with diabetes
corporate greed, lack of ethics sent us of the trajectory on generational diabetes.
if stories! dealing with diabetes burnout, you're not alone. read these you're
body mass index versus surrogate predictors of central adiposity as independent measures of mortality in type 2 diabetes

Técnica EDA com substituição de sinônimos
i shared something on fb about covid being vascular and it was word to a friend, who has diabetes, and has had covid.
research suggests vitamin d reduces blood sugar in people with diabetes
corporate greed, lack of ethical code sent us on the trajectory of generational diabetes.
if you're dealing with diabetes burnout, you're not solitary. read these stories!
body mass index versus surrogate quantify of central adiposity as independent predictors of mortality in type 2 diabetes

Fonte: Autoria própria (2023).

possível controlar a taxa de solicitações à API do Google. Os idiomas definidos foram inglês e espanhol, sendo o inglês o idioma original dos *tweets*.

No *back-translation* não é possível definir uma quantidade de substituições no texto, isso ocorre porque a API do Google tradutor apenas traduz a sentença para outro idioma e depois retorna para o idioma original, então não existe um controle do que será alterado. O Quadro 5 mostra cinco *tweets* sem DA, em seguida os mesmos *tweets* são exibidos com as técnica de *back-translation*. Ao observar as amostras das sentenças geradas com *back-translation* é possível afirmar, de modo geral, que essa técnica apresenta bons resultados na preservação do contexto.

Quadro 5 – Amostra de *tweets* para demonstração da técnica *back-translation*

Sem o uso de técnicas de DA
i shared something on fb about covid being vascular and it was news to a friend, who has diabetes, and has had covid.
research suggests vitamin d reduces blood sugar in folks with diabetes
corporate greed, lack of ethics sent us on the trajectory of generational diabetes.
if you're dealing with diabetes burnout, you're not alone. read these stories!
body mass index versus surrogate measures of central adiposity as independent predictors of mortality in type 2 diabetes
Técnica *back-translation*
i shared something in fb about covid was vascular and it was news for a friend, who has diabetes, and has had covid.
research suggests that vitamin d reduces blood sugar in people with diabetes
corporate greed, lack of ethics sent flight us on the trajectory of generational diabetes.
if you are dealing with the exhaustion of diabetes, you are not alone. these stories!
body mass index versus substitute measures of central adiposity as independent predictors of mortality in type 2 diabetes

Fonte: Autoria própria (2023).

5.3 Rotulagem do dataset

Nessa etapa, a rotulagem dos dados consiste em atribuir uma categorização aos dados textuais de *tweets*, com o objetivo de possibilitar o treinamento de modelos de ML.

Isso se faz necessário, pois no dataset obtido, os *tweets* não possuem rótulos. Tendo em vista apenas a aplicação das técnicas de DA ao dataset, não seria necessário rotular os *tweets*, porém a rotulagem é uma etapa importante para os modelos de classificação de texto definidos na metodologia, os quais serão utilizados para avaliar a eficácia das técnicas de DA. Dessa forma, a rotulagem foi aplicada tanto ao dataset original, quanto aos datasets gerados a partir das técnicas de DA.

O processo de rotulagem tem início com a geração de *embeddings* das sentenças dos *tweets* fazendo uso do modelo pré-treinado *stsb-distilroberta-base-v2*. A partir dos *embeddings*,

é possível que o algoritmo *k-means* forme agrupamentos de *tweets*, conforme a similaridade de contextos encontrada.

Nesse experimento, foram gerados seis *clusters* distintos com as seguintes características:

- ***Cluster 0***: *tweets* que em sua maioria tratam de pesquisas, métodos, profissionais e corporações que possuem algum envolvimento no tratamento ou cura da diabetes;
- ***Cluster 1***: consiste em *tweets* com pensamentos, sentimentos e depoimentos relacionados a diabetes na vida das pessoas;
- ***Cluster 2***: *tweets* que citam doenças como: câncer, depressão, lúpus, ansiedade e outras, relacionadas ou não com a diabetes;
- ***Cluster 3***: são *tweets* sobre obesidade, doenças cardiovasculares e maus hábitos alimentares;
- ***Cluster 4***: *tweets* que abordam vitaminas, frutas, ervas e demais alimentos benéficos para o diabético. Outro tópico comum nesse *cluster* está associado aos benefícios do exercício físico;
- ***Cluster 5***: maioria de *tweets* curtos, com pouco contexto e que não representam nenhum tópico específico.

O Quadro 6 mostra 2 exemplos de cada *cluster*. Esses exemplos foram comparados após o uso das técnicas de DA e foi constatado que as sentenças permaneceram no mesmo *cluster*. O número de *clusters* foi definido com base na representação visual obtida por meio do método t-SNE, que permitiu observar a distribuição dos *tweets* conforme a Figura 25. O processo é repetido com os mesmos parâmetros para as técnicas EDA na Figura 26 e *back-translation* na Figura 27.

Além disso, durante os experimentos, observou-se que ao repetir o processo de rotulagem para as técnicas de DA, os *clusters* formados apresentavam divergências significativas entre si. Essa divergência ocorreu devido à aleatoriedade do algoritmo k-means, que seleciona um ponto inicial aleatório para a formação dos *clusters*. Portanto, durante a geração dos *clusters* no conjunto de dados sem DA, o vetor de *centroids* foi salvo. Esse vetor contendo os *centroids* foi utilizado como parâmetro de inicialização para rotular os *tweets* no conjunto de dados aumentados com as técnicas de DA.

Sendo assim, foi possível observar que a base de dados aumentada continuou consistente e com as mesmas características da original, contudo, apresentou um volume maior.

Quadro 6 – Exemplos de *tweets* com seus respectivos *clusters* sem o uso de DA.

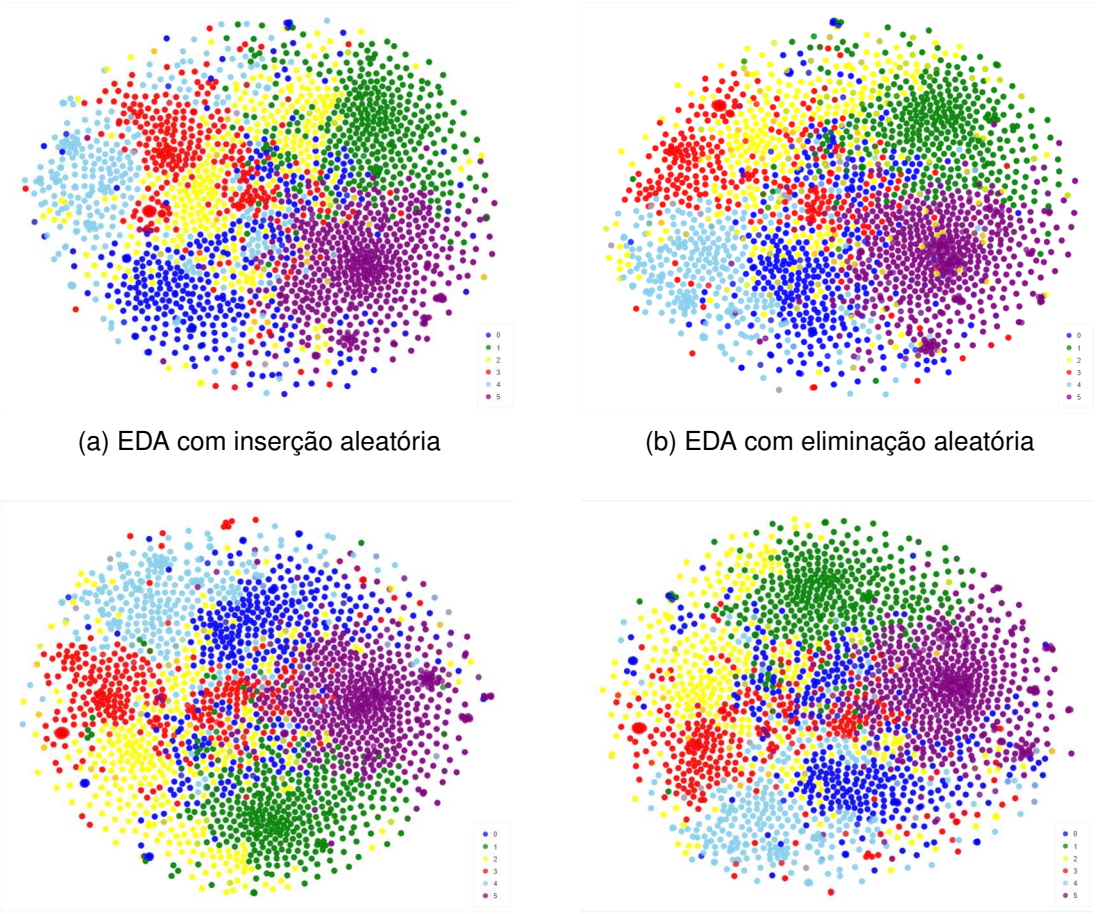
Clusters	***Tweets***
0	beautiful work delivering new molecular insight into type 2 iabetes treatment. congratulations to everyone involved!
0	scientists develop 12-hour method to predict diabetes onset in patients using artificialintelligence
1	i started eating candy to calm myself and he scored. i'm eating candy all match all tournament now even if it gives me diabetes.
1	i'm sorry you go through this friend. it never occurred to me until this morning that it could be diabetes related
2	2/6 childhood trauma can have a long-lasting impact on our health and well-being, leading to chronic illnesses such as, - type 2 diabetes - crohn's disease - anxiety - depression - and more
2	wouldn't that mean we end up like the us? hundreds for diabetes meds, thousands for surgery, absolutely no affordable mental health care.
3	venom, added sugar can cause diabetes, obesity and other diseases
3	obesity is a major health problem that the world is facing today and it is rising exponentially leading to other health complications like stroke, hypertension, diabetes, and even death.
4	vitamin d and risk for type 2 diabetes in people with prediabetes "in adults with prediabetes, vitamin d was effective in decreasing risk for diabetes."
4	eating fresh fruit can lower risk of diabetes according to a major new long-term study from oxford university
5	life with type 1 diabetes can be a struggle.
5	sweet diabetes

Fonte: Autoria própria (2023).

Figura 25 – Visualização dos *clusters* com t-SNE, sem DA.

Fonte: Autoria própria (2023).

Figura 26 – Visualização dos *clusters* com t-SNE, usando técnicas de EDA.



(a) EDA com inserção aleatória

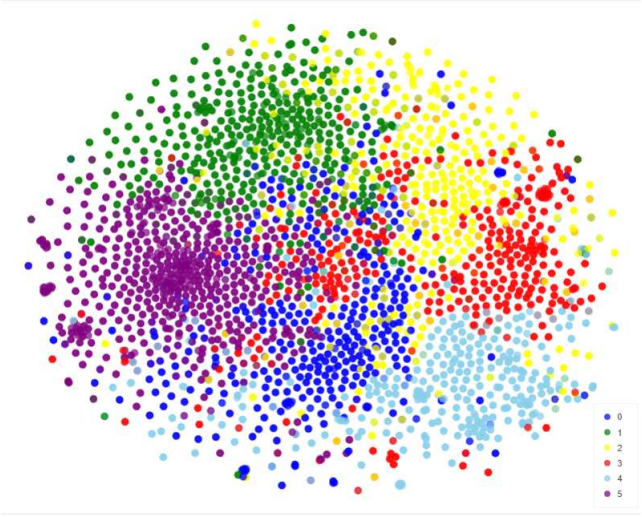
(b) EDA com eliminação aleatória

(c) EDA trocando aleatório

(d) EDA com substituição de sinônimos

Fonte: Autoria própria (2023).

Figura 27 – Visualização dos *clusters* com t-SNE na técnica de *back-translation.*



Fonte: Autoria própria (2023).

5.4 Resultados dos modelos de validação

5.4.1 Validação por métricas de desempenho quantitativas

Essa abordagem de validação utilizou três métodos de classificação - RF, SVM e NB -, objetivando tentar prever os rótulos dos *tweets*. Também foi realizada uma validação cruzada *k-fold* com 5 *folds* para obter a média de desempenho para a métrica de *accuracy*. O destaque nessa etapa do experimento foi a técnica de EDA com inserção aleatória e o método RF, que obteve uma precisão de **86%** na previsão dos rótulos. Dentre os modelos de classificação, o RF apresentou os melhores resultados em todas as técnicas de DA. Os demais resultados obtidos podem ser visualizados na Tabela 1.

Em um contexto geral, ambas as técnicas de DA apresentaram resultados superiores ao dataset sem DA para a classificação de texto. Portanto, presume-se que o uso das técnicas de DA pode ser validado no cenário em que se deseja aprimorar o desempenho dos modelos de classificação.

Tabela 1 – Precisão em (%) dos modelos de classificação

Técnicas de DA	**RF**	**SVM**	**NB**
Sem DA	54	56	51
EDA com inserção aleatória	**86**	**74**	**65**
EDA com eliminação aleatória	81	71	63
EDA trocando aleatório	84	73	64
EDA substituição de sinônimos	84	**74**	63
Back-translation	71	67	60

Fonte: Autoria própria (2023).

5.4.2 Validação por questionamento direto

Essa abordagem de validação consistiu no questionamento de 12 alunos do curso de Engenharia de Computação, sendo 3 deles egressos, os quais responderam à um questionário online com uma amostra de 10 *tweets* sem DA, comparando-os com os *tweets* gerados com as técnicas de DA. Não foi informado ao respondentes quais textos eram aumentados ou não.

Dentre os participantes, 8 alunos informaram possuir o nível de proficiência em inglês intermediário, 3 nível avançado e 1 básico. A finalidade dessa abordagem foi determinar se o significado das sentenças antes e depois do DA permaneceram os mesmos. Para facilitar o engajamento dos participantes para responderem ao questionário, foram selecionadas poucas amostras de *tweets*. Sendo assim, as amostras escolhidas representam apenas uma projeção dos resultados.

O Quadro 7 mostra a porcentagem de alunos que concordaram com a manutenção do significado das sentenças antes e depois das técnicas de DA. Ao analisar os gráficos gerados

com as respostas do questionário, observa-se que no conjunto de amostras com as técnicas de EDA a média de respostas positivas foi de 33,3% para inserção aleatória e eliminação aleatória, 25% em trocando aleatório e 45,8% na substituição de sinônimos. No conjunto de amostras com a técnica *back-translation* a média de respostas positivas foi de 50%, ou seja, em metade das amostras o significado das sentenças não foi alterado. Todavia, a alteração do significado das sentenças após o uso das técnicas de DA, não é significativa para alterar negativamente o desempenho dos modelos de classificação de texto.

Quadro 7 – Respostas do questionário

***Tweets* sem DA**	***Tweets* com DA**	**Técnica de DA**	**Respostas positivas**
Most of the people i see wearing masks should be hoping it prevents diabetes.	Most of the people i see wearing masks forbid should be hoping it prevents diabetes.	EDA inserção aleatória	8.3%
Are you the lady from the diabetes commercials?	Are you the lady dame from the diabetes commercials?		58.3%
Was your first clue that she didn't have diabetes?	Was your first clue that didn't have diabetes?	EDA eliminação aleatória	33.3%
The poor old man was afflicted with diabetes	The poor old man was diabetes		33.3%
Retinal cells may have the potential to protect themselves from diabetic retinopathy patient talk	Retinal cells may have the themselves to protect potential from diabetic retinopathy patient talk	EDA trocando aleatório	16.7%
I got diabetes from eating there	Eating got diabetes from i there		33.3%
Once again i am forced to remind people to avoid the diabetes if at all possible	Once again i am forced to cue people to avoid the diabetes if at all possible	EDA substituição de sinônimos	58.3%
Research suggests 'direct association' between nitrites from food additives and diabetes risk	Search suggests 'direct association' between nitrites from food additives and diabetes risk		33.3%
And the one person i know that has diabetes will never get off their insulin without a cure.	And the only person i know that he has diabetes will never leave his insulin without a cure.	*Back-translation*	50%
I don't often fat shame, but her whole family looks like the board of directors for the american diabetes foundation.	It is not usually the fat penalty, but his whole family resembles the board of directors of the american diabetes foundation.		50%

Fonte: Autoria própria (2023).

Ao considerar os resultados apresentados nesse capítulo, evidencia-se que o uso das técnicas de EDA e *back-translation* no corpus original de *tweets* contribuiu para melhorar o desempenho dos modelos de classificação de texto. Além disso, a análise das respostas dos alunos ofereceu uma percepção valiosa sobre a eficácia dessas técnicas na manutenção do significado das sentenças originais.

6 CONSIDERAÇÕES FINAIS

O objetivo principal desse trabalho foi validar a eficácia das técnicas de DA para textos, a fim de melhorar o desempenho de métodos de classificação textual. Para isso, foi realizada inicialmente uma pesquisa teórica sobre DA, o que foi indispensável para uma análise mais abrangente de suas abordagens em NLP, permitindo guiar este trabalho na linha de pesquisa proposta.

A obtenção de um corpus sobre diabetes foi um desafio considerável para o desenvolvimento deste trabalho. Primeiramente foi considerado utilizar registros médicos eletrônicos, porém são dados de difícil acesso devido à privacidade do paciente. A segunda opção foi utilizar a plataforma da mídia social Twitter para extrair um corpus sobre diabetes. Essa alternativa se mostrou viável apesar de algumas limitações da versão gratuita na API de desenvolvedor.

Definir as técnicas de DA foi um processo que envolveu muita pesquisa e leitura, entender o comportamento das técnicas e suas aplicações é um passo importante para a obtenção de bons resultados. Apesar disso, foram considerados outros aspectos para a escolha das técnicas, como o potencial em se adequar ao tema da proposta. Por exemplo, modelos de tradução como o *back-translation* têm a possibilidade de aumentar os dados, mesmo com dados escassos de idiomas pouco difundidos, além de fornecer um corpus de qualidade. A definição de quais técnicas seriam utilizadas também considerou pesquisas similares que utilizaram corpus extraídos do Twitter e o nível de complexidade de cada técnica.

Uma das etapas mais importantes do trabalho foi a validação, ou seja, verificar se os dados aumentados são funcionais e não foram deturpados durante o uso da técnica de DA. Sendo assim, para a validação foram propostas uma análise comparativa por meio de métricas de desempenho e uma validação manual por questionamento direto.

A validação por questionamento direto demonstrou que em algumas amostras o significado da sentença foi alterado após o DA. Em relação às métricas de desempenho, em todos os casos os dados aumentados com as técnicas de DA apresentaram resultados superiores a 60%. Apesar dos resultados serem animadores, se fazem necessários experimentos em maior escala, a fim de validar a abordagem de forma mais consistente.

Por fim, é possível concluir que apesar da melhoria no desempenho de tarefas de classificação de texto, as técnicas de DA escolhidas não apresentaram os resultados desejados no que diz respeito à preservação do significado das sentenças.

6.1 Trabalhos futuros

Em trabalhos futuros, sugere-se aprimorar o pré-processamento da base de dados para otimizar a qualidade do corpus. Realizar uma validação por questionamento direto com mais amostras e um número maior de participantes para melhorar a precisão dos resultados. Desenvolver funções próprias para o uso das técnicas de EDA, permitindo definir o parâmetro de

quantidade de ruídos, com verificação contextual a cada modificação na sentença. Essas contribuições futuras visam garantir uma preservação mais eficaz do contexto das sentenças, facilitar as análises e melhorar os resultados da classificação textual.

REFERÊNCIAS

ABULAISH, M.; SAH, A. K. A text data augmentation approach for improving the performance of cnn. *In*: IEEE. **2019 11th International Conference on Communication Systems & Networks (COMSNETS)**. [*S.l.*], 2019. p. 625–630.

AHMED, M.; SERAJ, R.; ISLAM, S. M. S. The k-means algorithm: A comprehensive survey and performance evaluation. **Electronics**, v. 9, n. 8, 2020. ISSN 2079-9292. Disponível em: https://www.mdpi.com/2079-9292/9/8/1295.

AMUTHADEVI, C.; UMAMAHESWARI, R.; BELINDA, M. Computational intelligence based medical diagnosis. **Journal of Computational and Theoretical Nanoscience**, American Scientific Publishers, v. 15, n. 6-7, p. 2369–2372, 2018.

AROYEHUN, S. T.; GELBUKH, A. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. *In*: **Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)**. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. p. 90–97. Disponível em: https://aclanthology.org/W18-4411.

BARRIERE, V.; BALAHUR, A. Improving sentiment analysis over non-english tweets using multilingual transformers and automatic translation for data-augmentation. **arXiv preprint arXiv:2010.03486**, 2020.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, p. 5–32, 10 2001.

CASELI, H.; NUNES, M. (Ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. [*S.l.*]: BPLN, 2023. https://brasileiraspln.com/livro-pln. ISBN 978-65-00-80693-9.

COULOMBE, C. Text data augmentation made simple by leveraging nlp cloud apis. **arXiv preprint arXiv:1812.04718**, 2018.

CULOTTA, A. Towards detecting influenza epidemics by analyzing twitter messages. *In*: **Proceedings of the First Workshop on Social Media Analytics**. New York, NY, USA: Association for Computing Machinery, 2010. (SOMA '10), p. 115–122. ISBN 9781450302173. Disponível em: https://doi.org/10.1145/1964858.1964874.

DAHL, A. A.; HALES, S. B.; TURNER-MCGRIEVY, G. M. Integrating social media into weight loss interventions. **Current Opinion in Psychology**, v. 9, p. 11–15, 2016. ISSN 2352-250X. Social media and applications to health behavior. Disponível em: https://www.sciencedirect.com/science/article/pii/S2352250X15002444.

DEAN, J. **Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners**. [*S.l.*]: Wiley, 2014.

DIABETES, S. B. de. Diretrizes da sociedade brasileira de diabetes 2019-2020. **Clannad Editora Científica**, 2019.

DU, J. *et al.* Self-training improves pre-training for natural language understanding. **arXiv preprint arXiv:2010.02194**, 2020.

FEDERATION, I. D. Idf diabetes atlas 10th ed. 2021.

FENG, S. Y. *et al.* A survey of data augmentation approaches for nlp. **ArXiv**, abs/2105.03075, 2021.

GARSIDE, R.; LEECH, G.; MCENERY, A. M. **Corpus Annotation: Linguistic Information from Computer Text Corpora**. [*S.l.*]: Routledge, 1997.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [*S.l.*]: MIT Press, 2016. http://www.deeplearningbook.org.

HARRIS, J. K. *et al.* Peer reviewed: Local health department use of twitter to disseminate diabetes information, united states. **Preventing chronic disease**, Centers for Disease Control and Prevention, v. 10, 2013.

HOU, Y. *et al.* Sequence-to-sequence data augmentation for dialogue language understanding. **arXiv preprint arXiv:1807.01554**, 2018.

IBRAHIM, M.; TORKI, M.; EL-MAKKY, N. AlexU-BackTranslation-TL at SemEval-2020 task 12: Improving offensive language detection using data augmentation and transfer learning. *In*: **Proceedings of the Fourteenth Workshop on Semantic Evaluation**. Barcelona (online): International Committee for Computational Linguistics, 2020. p. 1881–1890. Disponível em: https://aclanthology.org/2020.semeval-1.248.

JAIN, A.; KULKARNI, G.; SHAH, V. Natural language processing. **International Journal of Computer Sciences and Engineering**, v. 6, p. 161–167, 01 2018.

KANG, D. *et al.* Adventure: Adversarial training for textual entailment with knowledge-guided examples. **arXiv preprint arXiv:1805.04680**, 2018.

KARAMI, A. *et al.* Characterizing diabetes, diet, exercise, and obesity comments on twitter. **International Journal of Information Management**, Elsevier, v. 38, n. 1, p. 1–6, 2018.

KOBAYASHI, S. Contextual augmentation: Data augmentation by words with paradigmatic relations. **ArXiv**, abs/1805.06201, p. 452–457, 01 2018.

KUMAR, P. S.; UMATEJASWI, V. Diagnosing diabetes using data mining techniques. **International Journal of Scientific and Research Publications**, v. 7, n. 6, p. 705–709, 2017.

KUMAR, V.; CHOUDHARY, A.; CHO, E. Data augmentation using pre-trained transformer models. **arXiv preprint arXiv:2003.02245**, 2020.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, Nature Publishing Group UK London, v. 521, n. 7553, p. 436–444, 2015.

LI, B.; HOU, Y.; CHE, W. Data augmentation approaches in natural language processing: A survey. **AI Open**, Elsevier, 2022.

LI, Q. *et al.* **A Survey on Text Classification: From Shallow to Deep Learning**. 2021.

LIU, R. *et al.* Data boost: Text data augmentation through reinforcement learning guided conditional generation. **ArXiv**, abs/2012.02952, p. 9031–9041, 12 2020.

LIU, Y. *et al.* **RoBERTa: A Robustly Optimized BERT Pretraining Approach**. 2019.

LUQUE, F. M. **Atalaya at TASS 2019: Data Augmentation and Robust Embeddings for Sentiment Analysis**. 2019.

MAATEN, L. van der; HINTON, G. Visualizing data using t-sne. **Journal of Machine Learning Research**, v. 9, n. 86, p. 2579–2605, 2008. Disponível em: http://jmlr.org/papers/v9/vandermaaten08a.html.

MADHULATHA, T. S. An overview on clustering methods. **arXiv preprint arXiv:1205.1117**, 2012.

MARIVATE, V.; SEFARA, T. Improving short text classification through global augmentation methods. *In*: SPRINGER. **International Cross-Domain Conference for Machine Learning and Knowledge Extraction**. [*S.l.*], 2020. p. 385–399.

MCENERY, A.; XIAO, R.; TONO, Y. **Corpus-based language studies : an advanced resource book.** [*S.l.*]: Routledge, 2006. (Routledge Applied Linguistics Series).

MIAO, L.; LAST, M.; LITVAK, M. Twitter data augmentation for monitoring public opinion on covid-19 intervention measures. 2020.

MIKOŁAJCZYK, A.; GROCHOWSKI, M. Data augmentation for improving deep learning in image classification problem. *In*: IEEE. **2018 international interdisciplinary PhD workshop (IIPhDW)**. [*S.l.*], 2018. p. 117–122.

MILLER, G. A. Wordnet: a lexical database for english. **Communications of the ACM**, ACM New York, NY, USA, v. 38, n. 11, p. 39–41, 1995.

MITCHELL, T. M. **Machine Learning**. 1. ed. USA: McGraw-Hill, Inc., 1997. ISBN 0070428077.

N, T. R.; GUPTA, R. A survey on machine learning approaches and its techniques:. *In*: **2020 IEEE International Students' Conference on Electrical,Electronics and Computer Science (SCEECS)**. [*S.l.*: *s.n.*], 2020. p. 1–6.

N, T. R.; GUPTA, R. A survey on machine learning approaches and its techniques:. *In*: **2020 IEEE International Students' Conference on Electrical,Electronics and Computer Science (SCEECS)**. [*S.l.*: *s.n.*], 2020. p. 1–6.

OLIVEIRA, L. E. S. e *et al.* SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical NLP tasks. **Journal of Biomedical Semantics**, v. 13, 05 2022.

PAL, R.; PORAY, J.; SEN, M. Application of machine learning algorithms on diabetic retinopathy. *In*: IEEE. **2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)**. [*S.l.*], 2017. p. 2046–2051.

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

RAI, P.; SHUBHA, S. A survey of clustering techniques. **International Journal of Computer Applications**, v. 7, 10 2010.

REGINA, M.; MEYER, M.; GOUTAL, S. Text data augmentation: Towards better detection of spear-phishing emails. **arXiv preprint arXiv:2007.02033**, 2020.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *In*: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2019. Disponível em: http://arxiv.org/abs/1908.10084.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. **Journal of big data**, SpringerOpen, v. 6, n. 1, p. 1–48, 2019.

SUN, L. *et al.* Mixup-transformer: dynamic data augmentation for nlp tasks. **arXiv preprint arXiv:2010.02394**, 2020.

TURCHIN, A.; BUILES, L. F. F. Using natural language processing to measure and improve quality of diabetes care: a systematic review. **Journal of Diabetes Science and Technology**, SAGE Publications Sage CA: Los Angeles, CA, v. 15, n. 3, p. 553–560, 2021.

WANG, W. Y.; YANG, D. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. *In*: **Proceedings of the 2015 conference on empirical methods in natural language processing**. [*S.l.*: *s.n.*], 2015. p. 2557–2563.

WEI, J.; ZOU, K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. **arXiv preprint arXiv:1901.11196**, 2019.

WU, X. *et al.* Conditional bert contextual augmentation. *In*: SPRINGER. **International conference on computational science**. [*S.l.*], 2019. p. 84–95.

XIE, Q. *et al.* Unsupervised data augmentation for consistency training. **Advances in Neural Information Processing Systems**, v. 33, p. 6256–6268, 2020.

YU, A. W. *et al.* Qanet: Combining local convolution with global self-attention for reading comprehension. **ArXiv**, abs/1804.09541, 2018.

ZHANG, D. *et al.* On data augmentation for extreme multi-label classification. **arXiv preprint arXiv:2009.10778**, 2020.

ZHANG, X.; ZHAO, J.; LECUN, Y. Character-level convolutional networks for text classification. **Advances in neural information processing systems**, v. 28, 2015.

ANEXO A – Direitos Autorais - Lei N.º 9.610, de 19 de Fevereiro de 1998:

Disposições Preliminares

Presidência da República
Casa Civil
Subchefia para Assuntos Jurídicos

LEI Nº 9.610, DE 19 DE FEVEREIRO DE 1998.

Mensagem de veto

Vide Lei nº 12.853, de 2013 (Vigência)

Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências.

O PRESIDENTE DA REPÚBLICA Faço saber que o Congresso Nacional decreta e eu sanciono a seguinte Lei:

Título I

Disposições Preliminares

Art. 1º Esta Lei regula os direitos autorais, entendendo-se sob esta denominação os direitos de autor e os que lhes são conexos.

Art. 2º Os estrangeiros domiciliados no exterior gozarão da proteção assegurada nos acordos, convenções e tratados em vigor no Brasil.

Parágrafo único. Aplica-se o disposto nesta Lei aos nacionais ou pessoas domiciliadas em país que assegure aos brasileiros ou pessoas domiciliadas no Brasil a reciprocidade na proteção aos direitos autorais ou equivalentes.

Art. 3º Os direitos autorais reputam-se, para os efeitos legais, bens móveis.

Art. 4º Interpretam-se restritivamente os negócios jurídicos sobre os direitos autorais.

Art. 5º Para os efeitos desta Lei, considera-se:

I - publicação - o oferecimento de obra literária, artística ou científica ao conhecimento do público, com o consentimento do autor, ou de qualquer outro titular de direito de autor, por qualquer forma ou processo;

II - transmissão ou emissão - a difusão de sons ou de sons e imagens, por meio de ondas radioelétricas; sinais de satélite; fio, cabo ou outro condutor; meios óticos ou qualquer outro processo eletromagnético;

III - retransmissão - a emissão simultânea da transmissão de uma empresa por outra;

IV - distribuição - a colocação à disposição do público do original ou cópia de obras literárias, artísticas ou científicas, interpretações ou execuções fixadas e fonogramas, mediante a venda, locação ou qualquer outra forma de transferência de propriedade ou posse;

V - comunicação ao público - ato mediante o qual a obra é colocada ao alcance do público, por qualquer meio ou procedimento e que não consista na distribuição de exemplares;

VI - reprodução - a cópia de um ou vários exemplares de uma obra literária, artística ou científica ou de um fonograma, de qualquer forma tangível, incluindo qualquer armazenamento permanente ou temporário por meios eletrônicos ou qualquer outro meio de fixação que venha a ser desenvolvido;

VII - contrafação - a reprodução não autorizada;

VIII - obra:

a) em co-autoria - quando é criada em comum, por dois ou mais autores;

b) anônima - quando não se indica o nome do autor, por sua vontade ou por ser desconhecido;

c) pseudônima - quando o autor se oculta sob nome suposto;

d) inédita - a que não haja sido objeto de publicação;

e) póstuma - a que se publique após a morte do autor;

f) originária - a criação primígena;

g) derivada - a que, constituindo criação intelectual nova, resulta da transformação de obra originária;

h) coletiva - a criada por iniciativa, organização e responsabilidade de uma pessoa física ou jurídica, que a publica sob seu nome ou marca e que é constituída pela participação de diferentes autores, cujas contribuições se fundem numa criação autônoma;

i) audiovisual - a que resulta da fixação de imagens com ou sem som, que tenha a finalidade de criar, por meio de sua reprodução, a impressão de movimento, independentemente dos processos de sua captação, do suporte usado inicial ou posteriormente para fixá-lo, bem como dos meios utilizados para sua veiculação;

IX - fonograma - toda fixação de sons de uma execução ou interpretação ou de outros sons, ou de uma representação de sons que não seja uma fixação incluída em uma obra audiovisual;

X - editor - a pessoa física ou jurídica à qual se atribui o direito exclusivo de reprodução da obra e o dever de divulgá-la, nos limites previstos no contrato de edição;

XI - produtor - a pessoa física ou jurídica que toma a iniciativa e tem a responsabilidade econômica da primeira fixação do fonograma ou da obra audiovisual, qualquer que seja a natureza do suporte utilizado*;*

XII - radiodifusão - a transmissão sem fio, inclusive por satélites, de sons ou imagens e sons ou das representações desses, para recepção ao público e a transmissão de sinais codificados, quando os meios de decodificação sejam oferecidos ao público pelo organismo de radiodifusão ou com seu consentimento;

XIII - artistas intérpretes ou executantes - todos os atores, cantores, músicos, bailarinos ou outras pessoas que representem um papel, cantem, recitem, declamem, interpretem ou executem em qualquer forma obras literárias ou artísticas ou expressões do folclore.

XIV - titular originário - o autor de obra intelectual, o intérprete, o executante, o produtor fonográfico e as empresas de radiodifusão. (Incluído pela Lei nº 12.853, de 2013)

Art. 6º Não serão de domínio da União, dos Estados, do Distrito Federal ou dos Municípios as obras por eles simplesmente subvencionadas.