UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

# Geração de Sentenças para Classificação Semissupervisionada de Textos

**Ivan Caramello de Andrade**

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

**Ivan Caramello de Andrade**

# Geração de Sentenças para Classificação Semissupervisionada de Textos

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Ricardo Marcacini

**Versão original**

**São Carlos**

**2022**

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

C259g

Caramello de Andrade, Ivan
Geração de Sentenças para Classificação Semissupervisionada de Textos / Ivan Caramello de Andrade; orientador Ricardo Marcacini. -- São Carlos, 2022.
61 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. INTELIGÊNCIA ARTIFICIAL. 2. APRENDIZADO COMPUTACIONAL. 3. LINGUAGEM NATURAL. 4. REDES NEURAIS. I. Marcacini, Ricardo, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

**Ivan Caramello de Andrade**

# Sentence Generation for Semi-supervised Text Classification

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Ricardo Marcacini

**Original version**

**São Carlos**

**2022**

Folha de aprovação em conformidade
com o padrão definido
pela Unidade.
No presente modelo consta como
folhadeaprovacao.pdf

## AGRADECIMENTOS

Gostaria de agradecer aos professores e colegas da primeira turma do MBA IA e Big Data pelo apoio, colaboração e especialmente pelo conhecimento e aprendizado que pudemos cultivar juntos.

Agradeço também à Encora Brazil Division pelo tempo e dados disponibilizados para o desenvolvimento desse curso e pelo comprometimento com a educação e com o crescimento.

# RESUMO

CARAMELLO DE ANDRADE, I. **Geração de Sentenças para Classificação Semissupervisionada de Textos**. 2022. 65p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

Em muitos problemas relacionados a classificação automatizada de textos, a obtenção de grandes quantidades de exemplos rotulados é custosa ou impraticável, mesmo quando há um número significativo de exemplos disponíveis para treinamento. No entanto, com o avanço recente na área de modelos computacionais capazes de compreender e gerar linguagem natural, é possível utilizar esses modelos para a geração de dados sintéticos baseados tanto em dados rotulados como não-rotulados, permitindo que modelos mais robustos de classificação sejam treinados de maneria semissupervisionada. Enquanto técnicas mais rudimentares de geração de dados sintéticos se baseavam na permutação ou remoção aleatória de caracteres e palavras, modelos linguísticos modernos utilizam o contexto de uma sentença para aumentar o conjunto de dados de maneira mais realista. Técnicas baseadas em redes neurais com transformers, como a BERT, conseguem propor palavras que melhor preencheriam uma lacuna numa sentença, baseada no contexto como um todo, o que leva a textos sintéticos que mantém o valor semântico e o rótulo dos textos originais. Neste trabalho, algumas técnicas para aumento de dados através da geração sintética de exemplos rotulados foram exploradas, levando a resultados promissores sobre suas aplicações em domínios com grande restrição na disponibilidade de exemplos rotulados.

**Palavras-chave**: Inteligência Artificial. Aprendizado Semissupervisionado. Síntese de Textos. Aumento de Textos. Redes Neurais.

## ABSTRACT

CARAMELLO DE ANDRADE, I. **Sentence Generation for Semi-supervised Text Classification**. 2022. 65p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

In many problems related to automated text classification, obtaining large amounts of labeled examples is expensive or unfeasible, even when there are numerous unlabeled examples available for training. However, with the recent developments on computational models capable of natural language comprehension and generation, it is possible to use such models to generate synthetic texts based on both labeled and unlabeled data and allow more robust classification models to be trained in a semi-supervised way. While less advanced text synthesis techniques are based on the random permutation and remotion of characters and words, modern linguistic models can use a sentence's context in order to augment the dataset in a more realistic fashion. Techniques based on Transformer neural netowrks, such as BERT, are capable of suggesting words that would optimally fill an open space in a sentence, based on the full context, which leads to synthetic texts that keep the semantic value and labels of the original ones. In this monograph, some data augmentation techniques based on synthetic generation of labeled examples were explored, leading to positive results regarding their application on domains with major restrictions on labeled text availability.



**Keywords**: Artificial Intelligence. Semi-supervised Learning. Text Generation. Text Augmentation. Neural Networks.

## LISTA DE FIGURAS

## LISTA DE TABELAS

## SUMÁRIO

# 1 INTRODUÇÃO

## 1.1 Contextualização

Classificação semi-supervisionada é uma área do aprendizado de máquina que busca aprender padrões capazes de classificar entradas baseado num conjunto de treinamento que não possui rótulos em sua totalidade, em quantidade significativamente menor ao conjunto total dos dados (ZHU; GOLDBERG, 2009). Esse tipo de cenário é comum em diversas aplicações, especialmente quando há disponibilidade de dados, mas a avaliação e rotulação desses dados depende de atividade humana e é custosa ou, de maneira geral, limitada. Isso pode acontecer devido a falta de acesso aos rótulos (por exemplo num conjunto de respostas de questionário onde o rótulo era um atributo opcional) ou, mais comumente, devido ao custo em obtê-los (por exemplo, sendo necessário que um humano leia cada uma das entradas, as compreenda e classifique manualmente). É muito comum que a quantidade de dados sem rótulos seja ordens de grandeza maior do que os com rótulos, e a cardinalidade dos dados é essencial para que padrões complexos, como os envolvendo imagem ou texto, sejam aprendidos por um modelo de aprendizado de máquina.

Em cenários desse tipo, duas soluções iniciais podem ser identificadas: ignorar-se os rótulos e tratar o problema de maneira estritamente não-supervisionada; ou ignorar os dados não-rotulados e tratá-lo de maneira estritamente supervisionada. Ambas as abordagens acabam desprezando uma parcela muito relevante dos dados - no caso da abordagem não-supervisionada, o atributo que se quer inferir é perdido, o que dificulta muito a resolução de um problema direcionado, quando as classes são conhecidas de antemão; no caso da abordagem supervisionada, não só uma grande quantidade de dados é perdida (o que tende a diminuir a qualidade do modelo resultante) como não se pode garantir se os dados com rótulos são uma amostra representativa do problema como um todo, podendo levar a uma solução enviesada e de fraca generalização para problemas pertencentes ao domínio que não foi contemplado durante o treinamento.

Assim sendo, abordagens semi-supervisionadas são interessantes para que se possa utilizar todo o conjunto de dados disponíveis para se treinar um modelo de classificação, permitindo que as técnicas de aprendizado atinjam alta generalização (ao utilizar a grande cardinalidade de dados não rotulados) sem um viés excessivo sobre os dados rotulados e, ainda assim, obtenha um bom aprendizado, com boas taxas de precisão e acurácia sobre o atributo-rótulo sendo avaliado (ao utilizar os exemplos rotulados) (CHAPELLE; SCHOLKOPF; ZIEN, 2009).

Um dos domínios que mais sofre com a falta de exemplos rotulados é o de linguagem natural. Cada vez mais se tem disponíveis dados textuais das mais diversas naturezas,

como artigos científicos, notícias, mensagens instantâneas e mesmo logs de erros. Muitos desses textos são disponibilizados gratuitamente na internet, e podem ser acessados através de webcrawlers; ou são gerados e preservados automaticamente por sistemas de acompanhamento. Raramente esses textos são registrados com rótulos significativos, e a rotulação de linguagem natural exige compreensão de texto - uma tarefa que ainda apresenta limites quando feita de maneira algorítmica. Assim, é necessário encontrar uma maneira de utilizar um pequeno conjunto de textos rotulados por humanos junto do grande volume de texto não-rotulado e, com isso, poder treinar modelos que aprendam e reconheçam padrões numa gama de textos ampla e diversa. Em outras palavras, é necessário generalizar modelos para o universo de documentos textuais sem limitar-se ao viés de exemplos rotulados manualmente.

Embora a rotulação automatizada de textos não-rotulados seja uma tarefa difícil (e, inclusive, uma tarefa análoga à de classificação não-supervisionada), é possível utilizar a abordagem de Data Augmentation ("aumento de dados") para que novos dados artificiais sejam gerados (FENG *et al.*, 2021). Esses dados, por serem sintéticos, têm seus rótulos conhecidos, e podem ser utilizados para aumentar a quantidade de dados rotulados para treino. Apesar de aumentarem a diversidade de entradas, possivelmente aumentando a generalização e reduzindo o viés, é importante que os dados sintéticos sejam representativos do conjunto total de dados (incluindo os não-rotulados) para que essa melhoria em generalização se converta em aprendizado real. Caso seja possível aprender padrões não-supervisionados para geração de dados a partir do conjunto não-rotulado, a Data Augmentation será capaz de utilizar esse conjunto para melhorar o processo de treinamento supervisionado, permitindo um melhor aprendizado .

No caso específico de textos, a existência de uma quantidade muito grande de dados e também trabalhos na área, bem como a ocorrência natural de gramáticas e outros padrões de linguagem natural permitem que dados sintéticos sejam gerados utilizando conhecimento textual geral. Isso pode ser aprimorado ao utilizar-se de transferência de conhecimento, usando todo o corpo de padrões já conhecidos para gerar novas sentenças que sejam representativas do conjunto geral (evitando viés e melhorando a generalização) e, ao mesmo tempo, parametrizando essa geração para que ela apresente os rótulos desejados (melhorando os critérios objetivos de uma técnica supervisionada de aprendizado) (SHORTEN; KHOSHGOFTAAR; FURHT, 2021).

Considerando a importância da área de classificação de linguagem natural, a dificuldade na obtenção de textos rotulados e os avanços tanto na geração de aumento de dados para textos como na própria área de linguagem natural em si, existe um espaço a ser explorado na melhoria de técnicas de aumento para treinamento semissupervisionado de modelos de classificação de textos.

## 1.2 Justificativa e Motivação

Enquanto a área de classificação e reconhecimento de imagens conta com uma grande variedade de exemplos rotulados, a quantidade e variedade de textos rotulados é mais limitada - tanto pela grande variedade de rótulos possíveis como pela subjetividade intrínseca dos rótulos nesse contexto.

Nesse contexto, trabalhos de inteligência artificial voltados à linguagem natural têm investido em modelos auto-supervisionados, capazes de aprender padrões gerais de texto que podem, então, ser utilizados em outras aplicações através de transferência de conhecimento.

Ainda assim, a detecção de padrões gerais de textos, por si só, não é suficiente para atacar problemas que dependem de rotulação se tais rótulos não estiverem disponíveis. Essas mesmas técnicas podem, porém, ser utilizadas para a geração de rótulos, que por sua vez podem permitir o treinamento de soluções mais robustas - com menos viés e maior precisão - para problemas de linguagem natural.

Enquanto várias técnicas de Data Augmentation para linguagem natural são usadas comumente para adicionar variabilidade, a maior parte delas se baseia na modificação dos textos rotulados (através da substituição por sinônimos ou oclusão de palavras e caracteres, por exemplo), essas técnicas se limitam à estrutura dos dados originais e raramente permitem que o domínio seja ampliado de maneira significativa, enviesando o modelo ao subconjunto contemplado pelos exemplos.

Assim sendo, neste projeto propõe-se utilizar técnicas de síntese de texto para a geração de exemplos rotulados únicos, que introduzem maior variabilidade ao conjunto de testes. Essa abordagem pode permitir que modelos aprendam sobre um conjunto mais amplo de entradas e, assim, generalizem melhor as classes sendo aprendidas.

Tal como técnicas de Transfer Learning (TORREY; SHAVLIK, 2010) enriquecem o treinamento de um modelo ao adicionarem padrões reconhecidos sobre um conjunto muito maior de exemplos, um comportamento semelhante é esperado ao se utilizar Síntese de Sentenças para o treinamento de classificadores sobre textos.

## 1.3 Questões de Pesquisa e Objetivos

Este projeto tem, como objetivo principal, o desenvolvimento e aplicação de técnicas de data augmentation sobre textos de linguagem natural para treinar, de maneira semi-supervisionada, modelos de classificação de textos. Mais especificamente, espera-se comparar e estudar as aplicações da geração de sentenças para enriquecer conjuntos de treinamento de dados e expor classificadores a conjuntos de treinamento mais diversos.

O projeto inclui os seguintes objetivos a serem concluídos:

- Estudo, implementação e aplicação de técnicas de Data Augmentation para a geração de novos exemplos rotulados, na língua inglesa, mais especificamente: técnicas clássicas de Augmentation aleatória, como substituição por sinônimos e oclusão; etécnicas baseadas em síntese de sentenças.

- Treinamento e aplicação de um modelo de classificação de textos, utilizando-se as técnicas de Data Augmentation supracitadas durante o treinamento do modelo com o objetivo de se obter resultados mais robustos.

- Comparação de diferentes abordagens de Data Augmentation, bem como sua parametrização, variabilidade e robustez.

## 1.4 Metodologia

No desenvolvimento deste projeto será, inicialmente, feito um estudo sobre técnicas conhecidas na literatura sobre Data Augmentation de textos, bem como sobre técnicas de Síntese de Sentenças. Dentre essas técnicas serão selecionadas algumas abordagens de comparação, bem como técnicas de síntese de sentenças que podem ser aplicadas no problema de Data Augmentation.

Em específico, a biblioteca NLPAug (MA, 2019) será utilizada para prover ambos os tipos de Data Augmentation, utilizando-se do Augmenter "RandomAug" para aplicar as operações tradicionais e o Augmenter "ContextualWordEmbsForSentenceAug" para a síntese de sentenças. Esse segundo Augmenter pode utilizar vários modelos de predição de palavras e síntese de sentenças, como GPT-2 (RADFORD *et al.*, 2019) e XLNet (YANG *et al.*, 2019). Tanto o modelo pré-treinado como os hiperparâmetros do Augmenter serão estudados e ajustados.

Então, serão estudadas técnicas de classificação de textos de maneira supervisionada e suas diversas aplicações, selecionando-se um pequeno subconjunto de técnicas e problemas clássicos a título de comparação. Uma vez que o objetivo desse projeto é o estudo das técnicas para Data Augmentation, as técnicas de classificação não serão usadas tal qual encontradas na literatura, com a exceção do ajuste de eventuais hiperparâmetros.

As técnicas de síntese selecionadas serão aprimoradas para o objetivo em questão, estudando-se a influência de seus hiperparâmetros e sua capacidade de generalizar textos rotulados dos problemas selecionados.

Finalmente, as técnicas selecionadas - tanto de síntese quanto de classificação - serão utilizadas para se analisar o impacto da utilização de técnicas de síntese quando exemplos rotulados de maneira semi-supervisionada são adicionados ao conjunto de treinamento.

## 1.5 Visão geral do projeto

Nos próximos capítulos, uma análise dos trabalhos relacionados às áreas pertinentes - classificação semissupervisionada, representação computacional de textos e aumento de dados - serão explorados, seguidos pelo detalhamento da arquitetura que será utilizada no desenvolvimento deste trabalho.

Em seguida, os conjuntos de dados, técnicas e modelos utilizados serão detalhados, bem como detalhes e exemplos sobre a implementação das técnicas de aumento e representação textual. Serão apresentadas também as técnicas utilizadas para avaliação dos resultados finais.

Resultados experimentais serão, então, compartilhados e analisados, verificando o desempenho das técnicas propostas em diferentes cenários e parametrizações. Esse capitulo será seguido das análises finais e conclusões possiveis sobre o trabalho, bem como considerações sobre suas limitações e trabalhos futuros.

# 2 FUNDAMENTOS E TRABALHOS RELACIONADOS

Neste capítulo serão recapituladas as práticas mais comuns, bem como o estado da arte de três das principais áreas que embasam este trabalho: a classificação semissupervisionada, a representação computacional de textos e o aumento de dados. Em cada uma dessas sessões, diversas técnicas serão apresentadas e discutidas, com o objetivo de permitir a seleção das mais apropriadas para a resolução dos problemas e objetivos previamente declarados.

## 2.1 Classificação Semissupervisionada

Algoritmos de classificação são métodos de aprendizado de máquina que tem como objetivo principal atribuir rótulos a elementos considerando as características destes e as relações entre características e rótulos de outros elementos previamente vistos (ZAKI; JR, 2020). Essa abordagem normalmente depende de uma etapa de Treinamento, na qual são apresentados elementos, suas características e seus rótulos, de modo que as relações entre características e rótulos possam ser aprendidas através de uma técnica de aprendizado de máquina. Após a etapa de treinamento, é possível utilizar o modelo ou padrões extraídos da etapa anterior para estimar rótulos para novos elementos, realizando assim a classificação (AGGARWAL, 2015). No entanto, para muitos problemas do mundo real, a quantidade de dados rotulados é consideravelmente menor do que a totalidade deles (KWON *et al.*, 2020). Em outros casos, mesmo quando a quantidade de elementos rotulados é relativamente grande, não se pode garantir que eles são representativos. A falta de rotulação sistemática dos dados faz com que uma grande quantidade de elementos não possa ser utilizada para a etapa de treinamento em aplicações clássicas de classificação supervisionada, diminuindo a precisão dos modelos gerados e sua capacidade de representar o conjunto completo dos dados (ZHU; GOLDBERG, 2009).

Formalmente, os cenários de aprendizado de máquina podem ser divididos em supervisionado, não supervisionado e semissupervisionado. Num cenário supervisionado, um conjunto de n exemplos $X = (x_1, \ldots, x_n)$ é fornecido, assim como um conjunto de rótulos $Y = (y_1, \ldots, y_n)$, resultando num conjunto de pares de observações $S = (x_1, y_1), ..., (x_i, y_i)$. Num cenário semissupervisionado, por sua vez, o mesmo tipo de informação está disponível, mas apenas com um pequeno sub-conjunto de exemplos $X_l \subset X$ com seus respectivos rótulos $Y_l \subset Y$. Em contrate com o cenário supervisionado, a quantidade de exemplos não-rotulados $X_u \subset X$ pode ser consideravelmente grande. Finalmente, em aprendizado não-supervisionado a quantidade de conhecimento sobre os dados a serem analisados é ainda mais restritiva, com apenas um conjunto onde $n$ exemplos de $X$ sendo fornecidos, sem quaisquer rótulos $Y$.

Algoritmos de classificação baseados em aprendizado semissupervisionado são o foco deste projeto e têm como objetivo resolver problemas de classificação utilizando tanto dados não rotulados quanto dados rotulados (CHAPELLE; SCHOLKOPF; ZIEN, 2009). Essa utilização permitiria ao algoritmo gerar modelos que contemplam um conjunto maior de dados, potencialmente levando a uma melhor generalização e a melhores indicadores de métricas como precisão e acurácia. A utilização desses dados sem rotulação também permitiria a redução no custo de preparação dos dados, uma vez que a atividade de rotular dados, quando possível, costuma exigir análise humana e conhecimento da área. Nos problemas relacionados à linguagem natural, por exemplo, a rotulação costuma depender da leitura e interpretação de documentos textuais por humanos, uma atividade que exige recursos capacitados e tempo (AGGARWAL, 2018). Além disso, caso a exposição dos dados seja controlada - por exemplo, caso os dados provenham de informações sigilosas ou pessoais - a rotulação pode ser ainda mais limitada, com maiores restrições sobre como esses dados podem ser acessados e classificados (LIU *et al.*, 2021).

Apesar dessas diversas vantagens na utilização de dados não-rotulados junto com dados rotulados, estudos recentes demonstram que também é importante extrair características dos dados que possam ser utilizadas para melhorar o aprendizado supervisionado ou gerando instâncias rotuladas a partir dos rótulos disponíveis e dos exemplos não-rotulados (ENGELEN; HOOS, 2020). Uma das abordagens conhecidas para classificação semissupervisionada é a utilização dos exemplos rotulados para o treinamento de um modelo parcial (ou modelo base), que por sua vez é utilizado para a classificação dos exemplos não-rotulados, conhecida como *Self-Training* ('auto-treinamento') (YAROWSKY, 1995). A principal vantagem dessa abordagem é fazer uso de dados previamente não rotulados que, de outra maneira, estariam indisponíveis e não seriam utilizados pelo aprendizado. No entanto, essa abordagem implica na replicação de erro, pois classificações errôneas dos dados não-rotulados irão propagar esses erros, uma vez que os novos rótulos serão considerados como corretos quando o modelo geral, que leva em consideração todos os dados (rotulados e não rotulados) for treinado. Esse problema pode ser mitigado ao se utilizar apenas os exemplos não-rotulados cuja classificação foi realizada com alta probabilidade de acerto - diversas técnicas de classificação retornam valores de confiança, assertividade ou qualidade ao classificarem um elemento, e esses valores podem ser utilizados para a seleção dos dados a serem utilizados durante o treinamento, limitando o erro propagado (NING *et al.*, 2021). Essa etapa pode ser repetida diversas vezes, agregando dados rotulados nos modelos intermediários conforme o grau de confiança de seus rótulos for maior do que um valor de referência. Com isso, uma parcela maior dos dados pode ser utilizada, limitando-se a propagação máxima de erro em cada uma delas.

Mais recentemente, uma abordagem alternativa para apoiar o aprendizado semissupervisionado é a introdução de novos exemplos sintéticos, derivados de exemplos reais (LIU *et al.*, 2020), chamada de *data augmentation*. Nessa abordagem, ao invés de existir

uma etapa intermediária de treinamento para providenciar rótulos, há uma etapa que busca gerar novos exemplos rotulados baseados em exemplos gerais e em exemplos prévios rotulados. Dependendo do domínio, isso pode ser obtido através da obtenção de medidas estatísticas sobre os dados originais e, a partir delas, a adição de ruído proporcional à distribuição em exemplos reais. Essa abordagem adiciona variabilidade ao conjunto de dados sem propagar erros de um modelo intermediário, mas ela depende da fidelidade dos exemplos gerados pela adição de ruído. Esses ruídos não necessitam ser estritamente numéricos - a modificação de parâmetros como luminosidade ou orientação numa imagem ou a substituição de palavras por sinônimos são exemplos da aplicação dessa abordagem em outros domínios.

Em certas circunstâncias também é possível obter novos exemplos sintéticos a partir da combinação de exemplos reais. Essa combinação pode ser obtida a partir da concatenação, permutação ou 'cross-over' (isto é, combinação de partes de exemplos distintos), mas depende de um conhecimento sobre a natureza dos dados e pode não ser apropriada para todos os tipos de características. Em muitos casos, diversas propriedades de um exemplo são intrinsecamente ligadas à estrutura ou a certos atributos do mesmo, dificultando ou impossibilitando a geração de exemplos realistas com essas técnicas.

Mesmo quando não é possível combinar exemplos, a oclusão ou adição de ruídos estritamente sintéticos (isto é, que não condizem com a natureza original dos dados) pode ser útil para tornar um conjunto de dados mais robusto a erros e perturbações reais, como leituras errôneas ou faltantes que podem ocorrer num exemplo real (NISHI *et al.*, 2021). Essa técnica também pode auxiliar com generalização, uma vez que exemplos incompletos podem possuir informação suficiente para o aprendizado e detecção de padrões (que poderiam não ser identificados pelo modelo de aprendizado caso o mesmo obtivesse poucos erros através da análise de outros atributos).

Finalmente, a geração de exemplos sintéticos pode ser baseada não apenas na alteração de exemplos reais como também na geração de novos exemplos a partir de modelos generativos (FENG *et al.*, 2021). Essa abordagem exige o treinamento de um modelo capaz de gerar novos exemplos, uma tarefa muitas vezes tão ou mais custosa quanto o objetivo do treinamento original, mas caso o domínio do problema seja conhecido e amplo, é possível reutilizar modelos generativos para diversas aplicações. Por exemplo, no caso de textos pode-se utilizar um modelo generativo generalista para continuar sentenças e utilizar as sentenças ampliadas como novas entradas, mantendo os rótulos originais. Esses modelos generalistas também podem ser utilizados como modelos iniciais que, refinados com dados específicos do domínio, podem gerar dados sintéticos mais próximos dos reais sem a necessidade de se treinar um modelo inteiro sobre a aplicação (SHORTEN; KHOSHGOFTAAR; FURHT, 2021).

No contexto deste projeto, há interesse em investigar métodos de *data augmentation*

para dados textuais (FENG *et al.*, 2021; SHORTEN; KHOSHGOFTAAR; FURHT, 2021) e incorporá-los em métodos de classificação semissupervisionados. Para tal, nas próximas seções são apresentados e discutidos técnicas para representação de textos e uma descrição de métodos de *data augmentation* que serão analisados neste projeto.

## 2.2 Representação de Textos

A grande maioria das técnicas de aprendizado de máquina para dados textuais dependem da representação de textos em formatos estruturados (AGGARWAL, 2018). Enquanto diversos problemas têm dados em formatos próximos àqueles esperados por uma técnica computacional (como leituras de sensores ou dados estatísticos) ou que podem ser transformados em dados computacionais de maneira simples (como a transformação de atributos categóricos em conjuntos de atributos binários), nem todos os domínios tem a mesma facilidade de representação. Linguagem natural, áudio e imagem são exemplos de domínios nos quais a mera codificação dos dados de maneira direta não é suficiente, sozinha, para incluir também informações que dependem de contexto, tempo, posição e semântica (AGGARWAL; ZHAI, 2012). Assim sendo, sem uma representação adequada é virtualmente impossível que técnicas de aprendizado de máquina possam identificar padrões e classificar elementos nesses domínios sem que, antes, a informação seja codificada de uma maneira que inclua tais informações.

Enquanto as imagens ainda contam com a representação básica de *pixels* que, mesmo não contendo todo o contexto necessário, preservam certas medidas locais (como distância e similaridade local), a representação clássica de textos, com o uso de caracteres, não possui essa propriedade de forma direta, pois possui ambiguidade e é dependente da língua. A similaridade semântica pode mudar drasticamente com a modificação de um único caractere, por exemplo, e na maior parte dos contextos é necessário ler sequências deles (como palavras ou mesmo frases) para que relações entre textos (ou mesmo trechos de textos) possam ser estabelecidas. Dessa maneira, é necessário utilizar técnicas mais robustas de representação de texto.

Uma das técnicas mais clássicas para isso envolve modificar a unidade mínima de representação de caracteres para palavras. Assim, ao invés de considerar textos como sequências de caracteres, eles podem ser considerados como sequências de símbolos que têm significado mais representativo. Essa abordagem, conhecida como *Bag of Words* (AGGARWAL; ZHAI, 2012), representa cada palavra única num texto como um identificador, e o texto em si como um conjunto desses identificadores, preservando os significados. Com essa abordagem, é possível treinar-se modelos de classificação que possam aprender e reconhecer padrões baseados nos conjuntos de palavras.

Essa abordagem possui algumas limitações, sendo uma delas o tratamento de palavras derivadas ou relacionadas. Caso cada palavra seja registrada unicamente, conju-

gações ou variações de uma mesma palavra serão reconhecidas como palavras distintas. Isso tornaria o aprendizado e identificação de padrões mais desafiador, uma vez que a quantidade de identificadores únicos seria muito maior do que o necessário e, ao mesmo tempo, as relações entre muitos deles não seriam transmitidas, forçando o algoritmo de aprendizado a aprender que, por exemplo, “aprender”, “aprendia” e “aprendizado” são palavras relacionadas. Uma maneira de resolver esse problema é a aplicação de técnicas de pré-processamento que obtém morfemas (isto é, as unidades mínimas que carregam significado numa linguagem) de palavras ou radicais das palavras, os quais são usados para identificação. O processo de obtenção de morfemas pode, para várias linguagens, ser automatizado baseado num conjunto de regras estabelecidas por especialistas, mas também há técnicas não-supervisionadas para identificação dos mesmos. Essa abordagem também permite que palavras compostas por diversos morfemas possam ser registradas como relacionadas a diversos identificadores de uma vez, aumentando a capacidade de aprendizado e mesmo de generalização.

O segundo problema dessa abordagem é o tamanho do conjunto de identificadores. Uma vez que cada identificador único se tornaria uma característica dos dados (‘data feature’), um aumento nesse número poderia levar a uma maior dificuldade no aprendizado, dada a chamada ‘maldição da dimensionalidade’ (GEORGE; JOSEPH, 2014). Mais importante, enquanto o número de palavras ou morfemas únicos num texto de tamanho limitado também é limitado, a quantidade deles disponíveis numa língua natural pode ser consideravelmente grande, com milhares ou mesmo milhões de palavras e morfemas distintos. Um problema com essa dimensionalidade pode trazer dificuldades computacionais, lentidão durante o treinamento e alto custo até mesmo para representação.

Finalmente, essa abordagem, ao utilizar um conjunto não-ordenado, também perde o significado presente na ordem das palavras. Esse impacto depende da língua natural em questão, mas a grande maioria das línguas naturais modernas dependem de ordem para que o significado completo de um texto possa ser recuperado. Soluções para esse problema incluiriam a representação de um texto como uma sequência de identificadores (aumentando ainda mais a dimensionalidade, uma vez que cada posição da sequência pode conter qualquer um dos identificadores); ou a utilização de bigramas (ou n-gramas) - identificadores que marcam não palavras ou morfemas, mas sequências de 2 (ou n) identificadores, representando a ordem internamente em cada identificador (SIDOROV *et al.*, 2014). Ambas essas soluções dependem de um aumento muito grande da dimensionalidade sem, necessariamente, fazer um uso otimizado desse espaço dimensional.

Outras técnicas fazem um uso mais eficiente desse espaço dimensional ao representar não apenas a presença, unicidade ou contagem de palavras, mas também representar sua relação intrínseca com outras, seja utilizando estruturas semelhantes à apresentada anteriormente como base ou novas estruturas.

### 2.2.1 Word2Vec

Uma das técnicas utilizadas para uma melhor representação de linguagem natural, baseada nas estratégias anteriormente descritas, é o chamado *Word2Vec* (MIKOLOV *et al.*, 2013). Essa estratégia parte do conceito de identificar cada palavra ou morfema unicamente mas, ao invés de utilizá-los diretamente, essas características são projetadas em um espaço vetorial que, então, é treinado para representar não só todas as palavras do vocabulário original como também suas relações (como distância e similaridade) sobre um corpo textual.

Como o nome sugere, essa técnica representa cada palavra como vetor multidimensional, correspondente a uma posição nesse espaço. Relações entre palavras, então, podem ser calculadas ao se calcular operações entre esses vetores. A similaridade, por exemplo, pode ser dada pela similaridade de cosseno entre os vetores nesse espaço.

Para obter essa representação, é necessário treinar um modelo sobre um conjunto de textos, utilizando uma rede neural artificial simples. Por exemplo, essa rede recebe como entradas representações no formato de *bag of words* contínuas (isto é, uma lista de *bag of words* que representam uma sequência) ou n-gramas, e busca aprender, como saídas, as palavras que formam o contexto da palavra de entrada num dado texto. Essa técnica permite aprender, na camada escondida, as relações entre diferentes palavras, baseado na frequência e estrutura com que aparecem relacionadas. Outra arquitetura relacionada ao *word2vec* envolve apresentar como entrada o contexto e então prever qual é a palavra para tal contexto. Da mesma forma, a camada escondida contém uma representação latente das palavras denominada *word embedding*.

Como uma técnica baseada em redes neurais, a hiperparametrização da mesma tem alta influência na capacidade de generalização e aprendizado, tanto pelo tamanho da camada escondida (que se tornará o espaço vetorial onde as palavras são representadas) quanto o escopo do contexto considerado, que pode variar no número de palavras. O tamanho do contexto também deve levar em consideração a estratégia usada para obtenção dos morfemas e a língua natural alvo, uma vez que idiomas distintos podem necessitar de cadeias de tamanhos distintos para apresentar relações.

A técnica *Word2Vec* de representação também faz uso - ou, de maneira mais precisa, é usada através de - transferência de conhecimento ('transfer learning') (TORREY; SHAVLIK, 2010). Essa técnica, por sua vez, se dá pelo treinamento de uma rede e utilização não da sua estrutura fim-a-fim, mas especialmente das camadas escondidas como maneira de representar informação. Enquanto a *Word2Vec* como um todo aprende a completar ou continuar textos, sua camada escondida codifica uma representação generalizada para palavras num espaço vetorial. Essa representação, então, pode ser utilizada para outras aplicações como uma etapa de pré-processamento, transformando palavras e textos em

elementos que podem ser computados por outras técnicas de aprendizado.

Uma limitação do *word2vec* é que também desconsidera muitas propriedades relacionadas à ordem das palavras, bem como gera *word embeddings* estáticas, ou seja, a representação de uma palavra não é mais alterada depois do treinamento, mesmo na presença de outro contexto. Para mitigar tais limitações, foram propostos métodos para lidar com *word embeddings* dinâmicas, como o BERT, descrito a seguir.

### 2.2.2 BERT

Muitas abordagens relacionadas à linguagem natural tradicionalmente representam texto como uma sequência ordenada de palavras - de fato, a ordem em que as palavras se apresentam é importante para a compreensão do sentido, bem como a relação das palavras numa estrutura (PETERS *et al.*, 2018). Essa abordagem foi apoiada pelo desenvolvimento de técnicas de aprendizado de máquina que apresentam estruturas sequenciais direcionais, como redes neurais recorrentes (que permitem o aprendizado de séries temporais de dados) e outras técnicas (OTTER; MEDINA; KALITA, 2020).

Por muito tempo, uma das técnicas mais populares para processamento de texto como uma série temporal foram as redes redes *long short-term memory* (LSTM) (SUNDERMEYER; SCHLÜTER; NEY, 2012) ('memória de curto prazo longo', numa tradução livre), que permitem manter informações lidas anteriormente como um contexto que pode ou não ser esquecido, baseado em pesos e ativações, além do contexto imediato que é carregado de uma palavra para outra. Essa abordagem permitiu que textos fossem compreendidos para além da janela imediata de contexto em volta de uma dada palavra, mas ainda apresentam um problema: a natureza direcional dessa técnica faz com que, necessariamente, palavras sejam processadas na ordem em que são apresentadas. Mesmo que essa 'memória' possa ser carregada para palavras futuras, essa interação só ocorre uma vez que as palavras foram processadas. Apesar de intuitiva e análoga à maneira como humanos normalmente leem e falam, essa abordagem dificulta a compreensão (e, mais importante, a abstração do conhecimento) quando a informação-chave de uma sequência de palavras não se encontra em ordem mas a compreensão depende dessa ordem.

Uma tecnologia desenvolvida mais recentemente propõe uma solução alternativa para essa abordagem: a utilização de modelos capazes de auto-atenção chamados transformers ('transformadores') (VASWANI *et al.*, 2017). Esses modelos não dependem de uma leitura sequencial de dados e, ao invés disso, podem aprender a focar a "atenção" em *tokens* específicos de uma sequência. Isso permite, por exemplo, que frases inteiras sejam analisadas de uma vez, e a ordem em que as palavras são utilizadas para compreensão possa ser aprendida pelo próprio algoritmo. Essa abordagem se mostrou muito útil tanto para imagens quanto para textos, fornecendo independência da estrutura da língua e permitindo que contextos possam ser melhor compreendidos.

Dentre os modelos de aprendizado que fazem uso dessa técnica, o BERT (Bi-directional Encoder Representation from Transformers) (DEVLIN *et al.*, 2018) foi um dos primeiros e, até hoje, é um dos mais populares. Ele se baseia na ideia de construir a representação de palavras através de redes de *Transformers* sem, necessariamente, assumir uma ordem na apresentação das palavras. Isso é feito através da utilização de máscaras nas frases de treinamento: ao invés de treinar um modelo que 'prevê' a próxima palavra, o BERT busca descobrir a palavra ou conjunto de palavras faltante, permitindo que o método de atenção do *transformer* priorize a ordem e importância dos termos durante o aprendizado.

Assim como em outros modelos de NLP, o principal uso do BERT é a codificação que ele gera - isto é, a maneira como, nas camadas escondidas, as palavras são representadas. Isso permite que ele seja usado, através de *Transfer Learning*, como uma maneira de representar a relação entre palavras. Essa representação pode, então, ser usada como entrada para outros algoritmos de aprendizado ou mesmo outras aplicações estáticas.

Durante o treinamento, o BERT também utiliza a previsão de sentenças como um todo, permitindo não só que frases completas sejam geradas como também abrindo espaço para que a influência das palavras num próximo conjunto não dependa da posição das mesmas numa frase anterior. Essa estratégia é particularmente interessante por forçar o método a aprender associações tanto estruturais como semânticas e abstrair não só associações diretas (isto é, associações que surgiriam por palavras coexistirem na mesma frase) como também relações entre assuntos.

### 2.2.3 GPT

Enquanto outros modelos de linguagem natural partem do princípio de um objetivo supervisionado para obter uma representação que pode, então, ser generalizada para outros contextos através de *Transfer Learning*, a abordagem proposta pelo GPT (*Generative Pre-Training*) (RADFORD *et al.*, 2018) propõe um modelo originalmente não-supervisionado que não tem um objetivo específico, buscando apenas a geração de dados baseado num contexto. Essa abordagem, chamada de pré-treinamento generativo, serve então de entrada para outras aplicações (como classificação, geração para contextos específicos e análise) que é feita como aperfeiçoamento sobre o modelo pré-treinado.

Assim como o BERT, o GPT é baseado em Transformers que observam um contexto, mas com uma quantidade maior de camadas escondidas. No entanto, ao invés de utilizar a abordagem bi-direcional do BERT, a arquitetura do GPT apresenta uma estrutura mais profunda de *transformers* direcionais. Isso significa que ele preserva a abordagem 'clássica' de prever palavra por palavra, levando em consideração as anteriores - num contexto que pode variar de 16 a 512 *tokens*, dependendo do modelo treinado. Na versão clássica do GPT, esse pré-treinamento foi feito sobre a base de dados "BookCorpus", contendo 7000

livros sobre diversos assuntos - o suficiente para que o modelo aprendesse uma gama de contextos relativamente ampla e genérica.

Baseado nesse pré-treinamento, o modelo do GPT pode, então, ser calibrado com um treinamento supervisionado sobre um domínio específico (como um certo estilo de escrita ou tema), transferindo o conhecimento aprendido até então e enriquecendo sua capacidade preditiva. Essa abordagem conta com um custo adicional de treinamento para cada novo domínio ou aplicação específica que seja utilizada, mas o pré-treinamento original é aproveitado.

Uma segunda vantagem da arquitetura básica do GPT é sua capacidade de expansão. Dada a natureza direcional do modelo (que diminui o número de conexões) e sua capacidade em identificar e abstrair conceitos chaves com o mecanismo de auto-atenção, é possível expandir o modelo com um número maior de camadas *Transformer*. De fato, modelos sucessores ao GPT, como o GPT2 e GPT3, trouxeram um número muito maior de camadas, parâmetros e conjuntos de treinamento. O GPT-2 (RADFORD *et al.*, 2019), por exemplo, teve seu pré-treinamento feito sobre mais de 8 milhões de páginas de internet, ampliando a gama de contextos e conteúdos cobertos e que, somados a um aumento no número de camadas e parâmetros (chegando a 1.5 bilhões de parâmetros), permite gerar textos sobre conteúdos diversos e também formatos como diálogos, descrições e até versões rudimentares de programação.

Uma das versões mais recentes, chamado de GPT-3 (BROWN *et al.*, 2020), expande novamente esse conjunto ao lidar tanto com um conjunto de dados consideravelmente maior como com uma parametrização ainda mais robusta (contemplando mais de 100 bilhões de parâmetros entre suas diversas camadas de transformers), com um pré-treinamento feito sobre conjuntos como *CommonCrawl* (que contém milhões de páginas de internet em diversos idiomas e também inclui seus códigos fontes) e a totalidade da Wikipédia. Isso permitiu que o GPT-3 fosse capaz de realizar tarefas mais generalistas do que modelos anteriores, de escrita criativa a codificação de programas. Esse poder de generalização é baseado, majoritariamente, numa arquitetura que permite a manutenção de grandes quantidades de parâmetros e uma exposição a quantidades muito grandes de dados.

## 2.3 Data Augmentation

A prática de aumento de dados (*Data Augmentation* em inglês) se dá pela geração artificial de dados de entrada baseados em dados reais (FENG *et al.*, 2021). Ela é frequentemente utilizada em domínios nos quais é difícil obter dados reais, quando os dados reais disponíveis não são representativos das classes que se pretende cobrir, ou ainda quando a quantidade de dados rotulados, especificamente, é limitada.

Textos e linguagem natural como um todo são um domínio onde vários desses

problemas frequentemente ocorrem - muita vezes bases de dados textuais são restritas a alguns domínios, dependem de rotulação humana (que é custosa) ou simplesmente não cobrem corretamente um subconjunto do domínio que é importante para a pesquisa. Nesses casos, técnicas de *Data Augmentation* podem ser empregadas para gerar novos exemplos artificiais que permitam melhor treinar modelos baseados em aprendizado.

É importante ressaltar que dados sintéticos raramente são capazes de substituir dados reais, e que os dados sintéticos carregam limitações: dependendo da técnica utilizada, resultados sintéticos podem ter pouca variabilidade, serem uma representação pobre do conjunto total de dados ou gerarem resultados de pouca credibilidade (SHORTEN; KHOSHGOFTAAR; FURHT, 2021). É possível avaliar a qualidade de um dado sintético aumentado contra um conjunto de dados maior mas, como nos casos em que eles são aplicados é incomum ter grandes conjuntos de dados disponíveis para comparação, em muitos casos essa avaliação se dá sobre a comparação das métricas de modelos utilizando os dados originais contra os dados aumentados.

Especificamente em textos, a utilização de dados aumentados é particularmente comum quando os textos dependem de rótulos complexos (como baseados em opinião ou características que dependem de avaliação humana) ou quando lidam com informações sensíveis, como dados médicos ou de cunho pessoal. Nesses casos, o aumento pode ser usado não só para ampliar o número de dados disponíveis como também para anonimizar os dados originais, garantindo que nenhum dado sensível foi diretamente utilizado na composição dos modelos. Isso é particularmente importante no treinamento de modelos altamente complexos baseados em geração de textos, uma vez que dados sensíveis inseridos nas etapas de aprendizado podem ser replicados durante a execução do modelo (VAJJALA *et al.*, 2020).

Diversas técnicas de aumento de dados textuais existem na literatura, com diversos objetivos em mente. Neste trabalho, serão analisadas técnicas mais populares, conforme discutido em (MARIVATE; SEFARA, 2020; VAJJALA *et al.*, 2020; SHORTEN; KHOSHGOFTAAR; FURHT, 2021):

- Substituição, inclusão ou oclusão simples de caracteres. Essas técnicas são capazes de simular erros de digitação e entrada, e são úteis para adicionar robustez em modelos que receberão *input* direto de usuários, como *chatbots*. No entanto, para modelos que buscam executar classificação ou auxiliar na compreensão de textos essas técnicas não adicionaram variabilidade de conteúdos e, muitas vezes, tem uma contribuição pequena para o ganho de acurácia de um modelo.

- Substituição, adição ou oclusão de palavras. Ao contrário da técnica anterior, esta é capaz de modificar e expandir semanticamente o conjunto de dados, adicionando palavras que não estariam disponíveis anteriormente no corpus. No entanto, para

que esses dados sintéticos ajudem no aprendizado é necessário que as palavras modificadas não alterem por completo as características chave do texto que buscam ser aprendidas. Por exemplo, num sistema de análise de sentimento, substituir a palavra 'triste' por 'feliz' na frase 'eu estou triste' exigiria uma reinterpretação da mesma e mudança do rótulo, algo além do escopo da maior parte das ferramentas de substituição. Ao invés disso, técnicas de representação de texto - como *word2vec*, apresentado anteriormente - podem ser utilizadas para a substituição de palavras por outras que apresentem similaridade. No exemplo acima, 'triste' poderia ser substituído por 'chateado', adicionando variabilidade sem alteração do rótulo.

Para a adição de palavras é necessário utilizar um modelo que seja capaz de completar ou gerar sentenças, uma vez que a estrutura gramatical e semântica da frase precisa ser mantida. Assim sendo, essa variedade de aumento de dados textuais necessita do uso de uma das técnicas mais sofisticadas, como BERT e GPT apresentadas anteriormente. Por outro lado, essa abordagem é capaz de enriquecer consideravelmente os exemplos disponíveis para treinamento e indo além da simples substituição por sinônimos ou palavras semelhantes. Essa abordagem é capaz de adicionar adjetivos e advérbios, enriquecendo tanto o vocabulário quanto a estrutura e complexidade dos textos disponíveis.

- Extensão ou geração de novas sentenças. Assim como na técnica anterior, é necessário utilizar um modelo robusto capaz de gerar sentenças. Neste caso, porém, a adição de variabilidade é consideravelmente maior, adicionando não apenas vocabulário novo como também novas estruturas de frase, ampliando consideravelmente o conjunto de possibilidades. Por esse motivo, essa abordagem também é a que pode introduzir maior ruído e, potencialmente, levar a exemplos sintéticos mal rotulados, mal formados ou não representativos do conjunto original. Para se evitar esse problema é necessário configurar os hiperparâmetros disponíveis nos modelos generativos, como calor. Também é possível realizar o treinamento de *fine tuning* dos modelos generativos baseado no corpo textual disponível para treinamento. Essa abordagem, apesar de mais custosa (por exigir o treinamento de um modelo auxiliar adicional), é a que pode permitir o maior enriquecimento e aumento dos dados disponíveis.

Considerando que as técnicas acima citadas podem ser utilizadas para gerar novos dados sintéticos que preservam diversas características dos dados originais, adicionando variedade ao conjunto de textos sem, necessariamente, alterar características de alto nível como o assunto ou o sentimento geral do mesmo, é possível assumir não só que esses dados podem fazer parte do mesmo conjunto $X$ de exemplos utilizados para gerá-los, como também que eles preservam os mesmos rótulos $Y$ dos exemplos originais. Essa abordagem pode, então, ser utilizada como aprendizado semissupervisionado: apesar de existir apenas um subconjunto $X_l \subset X$ com rótulos $Y_l \subset Y$, um novo conjunto $Xsynth$ pode ser gerado

a partir de $X$ (utilizado para treinar ou refinar os modelos geradores) e $X_l$ (utilizado como entrada para os modelos utilizados no processo de *data augmentation*), assumindo os rótulos $Y_l$ como rótulos de $Xsynth$, $Ysynth$. Apesar dessa suposição trazer possíveis erros vindos da geração de $Xsynth$, o aumento da quantidade de dados e, especialmente, o aumento da variedade e representatividade dos mesmos (ao utilizar $X$ como um todo, e não apenas $X_l$) podem levar a modelos mais robustos e representativos do que se apenas $X_l$ fosse utilizado.

# 3 DESENVOLVIMENTO

Neste capítulo, detalhes sobre as técnicas selecionadas para a realização de experimentos - tanto preliminares como finais - serão discutidos e apresentados. A arquitetura da solução e os conjuntos de dados utilizados para treinamento e avaliação também serão trazidos, bem como as parametrizações utilizadas para obter os resultados no capitulo seguinte.

## 3.1 Considerações Iniciais

Baseado nos princípios levantados anteriormente, o objetivo deste trabalho é realizar o treinamento de um modelo de classificação supervisionado utilizando dados sintéticos cuja geração foi obtida a partir de um sistema semi-supervisionado, com um conjunto de dados não rotulados sendo usado para treinar um modelo preditivo de linguagem natural e, então, um pequeno conjunto de exemplos rotulados sendo modificado para dar origem a novos exemplos rotulados que agregam representatividade e diversidade ao treinamento.

Para obter tais resultados, será necessário utilizar um conjunto de modelos textuais pre-treinados com uma boa representação da língua inglesa e que também permitam o *finetuning* das camadas intermediarias/escondidas. Isso é necessário pois, uma vez que a maior parte dos dados não terão rótulos conhecidos, é necessário que sua informação seja aprendida nos próprios *embeddings* das palavras. Em outras palavras, espera-se que o modelo textual aprenda a relação entre as palavras presentes no conjunto não-rotulado de treinamento, mesmo sem saber diretamente a relação das mesmas com rótulos. Isso permitirá que palavras relacionadas apareçam nos textos sintéticos mesmo que nunca ocorram no conjunto original de dados rotulados.

Além disso, também será necessário utilizar técnicas de *data augmentation* baseadas no modelo textual citado acima, bem como o uso de um modelo de classificação de textos, o qual será utilizado para o treinamento do modelo final. Neste trabalho, o modelo de classificação em si é indiferente, uma vez que as melhorias propostas ocorrem no enriquecimento e produção dos dados utilizados para treinamento. De maneira similar, as técnicas propostas são independentes do objetivo de classificação, que depende do conjunto de dados sendo utilizado.

As principais etapas deste trabalho podem ser vistas no diagrama 1, "Etapas da classificação semi-supervisionada de textos através de data augmentation". Em azul estão apresentados os dados rotulados; em vermelho, dados não rotulados; em amarelo, etapas relativas ao pre-processamento semi- e não-supervisionado; em verde, etapas relativas a modelos supervisionados de classificação; e, em roxo, etapas relativas a validação.

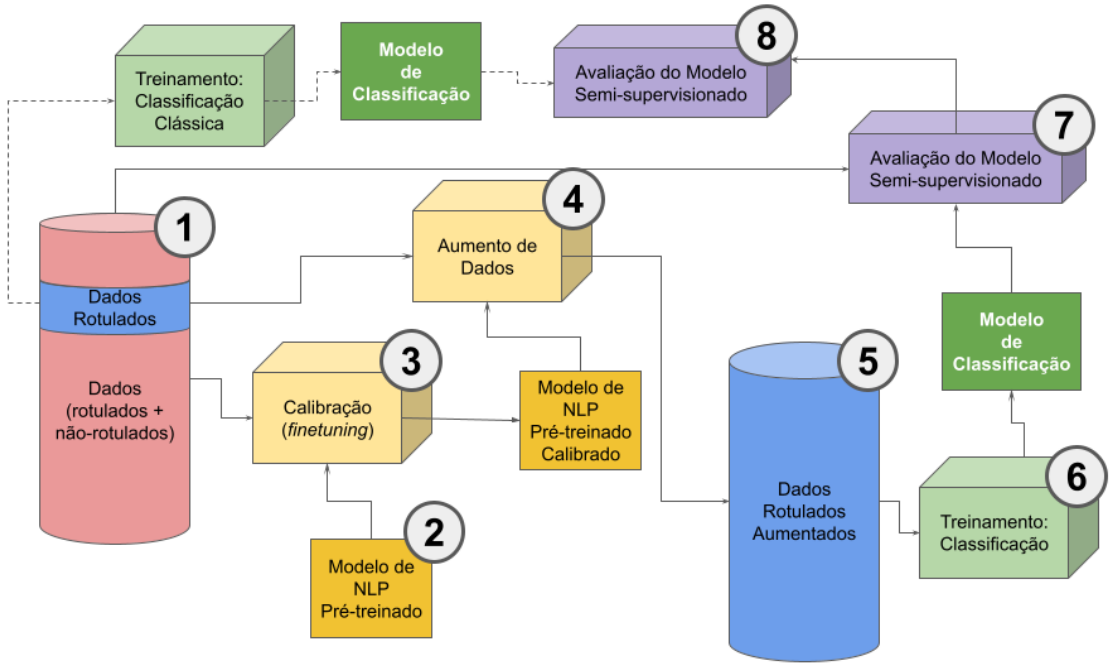


Figura 1 – Etapas da classificação semi-supervisionada de textos através de data augmentation

Cada uma das etapas é detalhada a seguir:

1. Obtenção de um conjunto de dados parcialmente rotulados, dentre os quais uma parcela minoritária dos dados tem rótulos conhecidos, e divisão deste conjunto em treinamento e teste (atentando-se para uma representação equilibrada de classes entre os conjuntos).

2. Obtenção de um modelo textual robusto pre-treinado na língua inglesa. Ao utilizar um modelo pre-treinado como entrada, dados não-rotulados (ou implicitamente rotulados), de fora do conjunto disponível para treinamento, automaticamente passam a compor o conjunto utilizado para gerar e comparar sentenças, adicionando mais uma camada semi-supervisionada ao projeto e, mais importante, permitindo que esses modelos representem relações mais ricas e completas do que as disponíveis apenas no conjunto de treinamento.

3. Calibração (finetuning) do modelo textual pre-treinado com os dados de treinamento completos (incluindo não-rotulados). Como o finetuning de um modelo pode ser feito nas camadas escondidas de embedding, treinando-se por exemplo o aprendizado de sentenças relacionadas, sinônimas ou sequenciais, é possível utilizar todo o dataset não-rotulado para a melhoria do modelo.

4. Aumento de dados através da geração de sentenças sintéticas a partir do modelo textual pre-treinado e das sentenças de treinamento rotuladas. Essa etapa permite

combinar o conjunto de treinamento não-rotulado com o rotulado, caracterizando uma etapa semi-supervisionada para a geração de exemplos sintéticos. Nessa etapa, os exemplos rotulados serão aumentados com técnicas de data augmentation que utilizam dados não-rotulados no seu treinamento, permitindo que as novas sentenças, ao mesmo tempo, mantenham a classe das sentenças originais e apresentem a variedade de vocabulários e formas presentes no conjunto original, mantendo também coesão sintática, semântica e gramatical graças ao modelo pre-treinado usado originalmente.

5. Divisão dos dados sintéticos em conjuntos de treinamento e testes, utilizando o rótulo dos dados utilizados na geração como rótulo dos dados sintéticos.

6. Treinamento de um modelo de classificação de textos baseado nos dados de treinamento sintéticos obtidos anteriormente. Essa etapa também pode ser baseada num modelo textual que tenha sido pre-treinado com o conjunto de dados não-rotulados, enriquecendo a capacidade de classificação e utilizando uma gama mais diversa de exemplos.

7. Avaliação do modelo de classificação obtido a partir dos exemplos sintéticos, calculando-se medidas como precisão, acurácia, recall e f1-score contra os dados de teste originais (não-sintéticos).

8. Comparação dos resultados obtidos contra o treinamento de modelos de classificação (similares aos da etapa de classificação) utilizando apenas os exemplos rotulados originais, sem dados sintéticos ou aumento de dados. Essa etapa serve para demonstrar se os resultados finais são aprimorados pela utilização de exemplos sintéticos enriquecidos com dados não-rotulados. Com o objetivo de verificar o efeito desse uso no cenário mais comum, ie, onde não existe um grande conjunto de dados rotulados, uma parcela pequena dos dados rotulados será considerada para os experimentos.

## 3.2 Datasets e preparação dos dados

Esse trabalho utiliza três conjuntos de dados disponíveis publicamente: um *dataset* primário, que contem sentenças e rótulos a serem utilizadas durante o treinamento, teste e validação das técnicas empregadas; um *dataset* externo, utilizado para realizar o *finetuning* do modelo de geração de sentenças; e um terceiro *dataset*, utilizado para comparação dos resultados em outro contexto.

O *dataset* primário utilizado nesse projeto é um *dataset* conhecido que contem *tweets* sobre o tema Covid-19. Esse *dataset* contem mais de 40000 entradas, classificadas por sentimento, de 'muito negativo' a 'muito positivo'. Por ser um dataset conhecido e bastante utilizado, ele será usado como *baseline* de comparação: serão treinadas redes de classificação utilizando esse *dataset* tanto na íntegra como com parcelas limitadas de seus

rótulos, com o intuito de se verificar se a utilização de dados sintéticos pode substituir ou melhorar os resultados quando poucos rótulos estão *disponíveis*.

Um *dataset* externo de sentenças será utilizado também para executar a etapa de *finetuning*. Esse *dataset* advém da Wikipedia, e é similar àquele utilizado original para treinar o modelo BERT. Ele será utilizado para diferenciar as frases do *dataset* primário de outras, permitindo que termos não-relacionados ao assunto em questão sejam penalizados durante a etapa de *data augmentation*.

Além disso, um terceiro *dataset*, contendo transcrições de chamadas telefônicas realizadas por robôs, foi considerado para a exploração qualitativa da técnica perante a análise de especialistas no dominio, a titulo de verificar se as sentenças geradas seriam capazes de manter e emular rótulos. Esse dataset não será utilizado nos resultados numéricos do capítulo seguinte, uma vez que sua utilização foi feita num contexto meramente qualitativo, sem medidas especificas de acurácia.

Uma vez selecionados os *datasets*, o *dataset* primário foi limitado a contar apenas as entradas com sentimentos 'muito negativo' e 'muito positivo', com o intuito de tornar o problema binário e, assim, facilitar a visualização e análise de precisão obtida.

Então, o *dataset* primário foi amostrado: uma pequena parcela (variando de 20% a 1%) foi mantida com os rótulos, enquanto o resto dos dados teve seus rótulos removidos para treinamento. Essa etapa representa a obtenção de um *dataset* com dados parcialmente rotulados.

Finalmente, uma parcela dos dados originais será reservada para validação. O restante dos dados será utilizado, num primeiro momento, para pre-processamento não supervisionado e finetuning dos modelos e, depois, os dados sintéticos resultantes serão divididos em conjuntos de treino e teste (preservando-se apenas os dados reais para validação).

## 3.3 Técnicas Utilizadas

Na execução deste projeto, dois conjuntos principais de técnicas foram utilizadas:

- Técnicas de processamento de linguagem natural, que incluem:
  - Tokenização de textos
  - *Embeddings*
  - *Fine tuning*
  - Classificação
- Técnicas de aumento de dados, que incluem:

- Substituição de palavras
- Adição de palavras

### 3.3.1 Técnicas de Processamento de Linguagem Natural

Na primeira categoria, optou-se por utilizar o modelo BERT, mais especificamente o modelo sem diferenciação de maiúsculas. Esse modelo pré-treinado conta com 110M parametros e 12 camadas *transformer*, e foi pre-treinado utilizando os conjuntos de dados Wikipedia e Bookcorpus com 15% de mascaramento por sentença para treinamento.

Como explicado no capitulo anterior, esses modelos podem ser utilizados tanto para *embedding*, criando representações intermediárias dos textos, como também para classificação de palavras ao se adicionar camadas completamente conectadas no final da rede. Assim sendo, as camadas intermediárias de *embedding* serão inicialmente utilizadas com base no modelo pre-treinado e, então, sofrerão *finetuning*. O modelo com *finetuning*, mas sem camadas finais de classificação, será utilizado para a geração de sentenças, enquanto as camadas finais serão treinadas com os dados sintéticos.

Devido a diferenças nos diferentes arcabouços utilizados pelas bibliotecas utilizadas, o modelo BERT será refinado utilizando-se a implementação em PyTorch da seguinte maneira: pares de sentenças sequenciais do *dataset* de treinamento serão utilizadas com um rótulo positivo, indicando que tem alta relação; e, então, pares compostos por sentenças do *dataset* de treinamento e sentenças de um *dataset* de apoio similar ao original utilizado para treinar o BERT serão utilizadas com um rótulo negativo, indicando que não tem relação. Isso será feito com as camadas escondidas descongeladas, de modo que esse conhecimento possa ser propagado para as camadas intermediárias. Esse passo será feito com apenas duas épocas, com o objetivo de minimizar o retreino dos pesos do modelo pré-treinado original. Os modelos de classificação farão uso de um número maior de épocas de treinamento.

### 3.3.2 Técnicas de Aumento de Dados

Para o aumento de dados, foi utilizada a biblioteca NLPAug, que fornece diversas técnicas de aumento de dados. Para este trabalho, priorizou-se as técnicas de aumento de sentenças, ou seja, técnicas que modificam o texto baseado no contexto de uma sentença como um todo, e não apenas no nível de palavras ou caracteres.

O modelo BERT pre-treinado e com *finetuning* semi-supervisionado foi utilizado em conjunto com a biblioteca NLPAug, de modo que os mesmos *embeddings* seriam utilizados para geração das sentenças sintéticas e sua posterior classificação.

As técnicas usadas foram a Inserção de palavras e a Substituição de palavras. Ambas escolhem, aleatoriamente, um numero de palavras a serem adicionadas ou substituídas numa sentença de entrada (limitado por uma parametrização), e então utilizam o modelo

| Texto Original | Texto Aumentado |
|---|---|
| I had some N95 masks from when we here in the SF Bay Area had to deal with smoke from the fires up north. I gave away all of them except for one. Will wearing it in the grocery store, pharmacy, etc. help prevent infection? Thanks for | i had some n95 stuff from when we here in the sf bay area had to deal with things from the storm up north. i gave away all of them thanks for one. will store it in the grocery store, pharmacy, etc. help solve that? look for |
| I am 100% right with God and ready. Are you? I don't make the rules, God does. Turn to Jesus immediately.#Jesus #JesusSaves #RepentNowNations #coronavirus #CoronavirusUSA #GodWins #panicbuying #toiletpaper #marketcrash2020 #TodayIsTheDay #COVID19 #Covid_19india #Quarantine | i am 100 % right with god and peter. are you? i don't make the decisions, god does. turn to jesus v. # jesus # nazareth # repentnownations # daniel # joseph # godwins # john # jesus # jesus # todayistheday # covid19 # covid _ 19india # quarantine |
| Lots of tweets reminding us if we social distance we save lives & if we're cavalier we cause deaths. Action over #COVID2019 got delayed for the sake of stock prices. | lots of tweets call us if we social responsibility we save lives & amp ; if we're responsible we earn deaths. action over # covid2019 got delayed for the sake of raise prices. |
| How quickly the world changes. Meanwhile actors and pro athletes arenÂt working and practicing. The world changed. Thank you health care workers, grocery store employees and truck drivers | how quickly the world transforms. meanwhile actors and pro engineers... studying and practicing. the world. thank you health care executives, shoe store employees and truck mechanics. |

Tabela 1 – Exemplos de *tweets* originais e versões aumentadas

de geração de textos para selecionar uma palavra dentre as mais prováveis para ser, respectivamente, inserida ou substituída no lugar de outra. O conjunto de palavras mais prováveis é obtido pelo próprio modelo linguístico, enquanto o numero de palavras consideradas é parametrizado. Essa etapa pode ser repetida diversas vezes numa mesma sentença, levando a alterações mais drásticas.

Experimentalmente, verificou-se que a inserção de palavras tende a trazer mais ruido e gerar sentenças com menor grau de realismo, enquanto a substituição de palavras tem um comportamento mais próximo do esperado para este projeto.

Exemplos das frases aumentadas podem ser vistos na tabela 1.

## 3.4 Implementação

Para cada uma das etapas especificadas, conjuntos de parâmetros foram utilizados para que diferentes alternativas fossem estudadas. Esses conjuntos de parâmetros foram selecionados em testes preliminares e, então, aplicados nos experimentos finais.

Para cada uma das etapas, os parâmetros disponíveis foram:

- Obtenção do conjunto de dados primário:
  - Separação de treinamento e teste: 80% para treino, 20% para teste
  - Proporção de rótulos mantidos: 20% e 5%
- Modelo pré-treinado: "BERT uncased"
- *Finetuning pré-aumento (se aplicado)*:
  - Dados: Combinação de 10000 sentenças do conjunto de treinamento primário entre si e de sentenças do conjunto de treinamento primário com sentenças do conjunto de dados externo.
  - Treinamento: duas épocas com taxa de aprendizado 0.0002, com todas as camadas configuradas para treinamento.
- Geração de sentenças sintéticas via aumento de dados:
  - Técnicas aplicadas:
    - Substituição;
    - de 5 a 25 palavras por sentença
    - seleção dentre as [5, 10, 15, 20, 25, 30, 40, 50, 75, 100, 125, 150, 200] palavras mais prováveis
  - Total de sentenças aumentadas:
    - 1 por sentença de entrada
    - 2 por sentença de entrada
    - 5 por sentença de entrada
    - 10 por sentença de entrada
- Treinamento de modelos de classificação:
  - Modelo utilizado: "BERT uncased"pre-treinado e com *finetuning*
  - *Finetuning*: apresentação dos dados não-rotulados por uma época.
  - Épocas: de 5 a 20
  - Taxa de aprendizado: de 0.0001 a 0.0002

## 3.5 Análise e Validação

A validação dos resultados é feita a partir da extração das métricas de avaliação contra os conjuntos de teste, e então sua comparação contra os modelos treinados sem os dados sintéticos. Foram também treinados modelos com os dados com todos os rótulos disponíveis, de modo a estabelecer um *baseline* de resultados ótimos possíveis.

Em todos os casos, os mesmos exemplos de validação foram utilizados - exemplos sintéticos não foram utilizados nessa etapa, uma vez que sua verossimilhança não foi testada ou validada. Ainda assim, é importante notar que os resultados obtidos só podem ser considerados como uma métrica de qualidade do aprendizado baseado em dados sintéticos por haver um conjunto de testes que não só apresenta rótulos como também é representativo do conjunto de dados como um todo. Em exemplos reais, onde o conjunto de dados rotulados pode apresentar viés de amostragem, a comparação final pode estar enviesada na direção dessa amostragem. Como o intuito da técnica proposta é adicionar representatividade e variedade aos dados, é possível que os resultados sobre um subconjunto das entradas seja pior do que sobre um conjunto que apresenta a mesma variedade do conjunto completo.

Para esses experimentos, as métricas de Acurácia, Precisão, *Recall* e o f1-score foram utilizados para comparação, bem como o tempo total de treinamento e pre-processamento. No entanto, uma vez que os rótulos se apresentam balanceados e possuem igual valor, o critério de Acurácia foi utilizado como principal. Os valores foram comparados entre diferentes parametrizações da técnica semissupervisionada e diferentes proporções do conjunto de treinamento com rótulos. Para cada uma delas, os resultados também foram comparados com os valores obtidos pelos modelos de classificação sem dados sintéticos, bem como com os valores obtidos pelos modelos que utilizaram os dados rotulados em sua completude. Os primeiros serviriam como a comparação principal, sendo o cenário no qual se pretende atuar, enquanto o segundo seria utilizado para mostrar o quanto seria possível aprender do conjunto de dados se todos os rótulos estivessem disponíveis.

# 4 RESULTADOS PRELIMINARES

Inicialmente, foram realizados testes preliminares que mostraram que a utilização da técnica proposta aumentou consideravelmente as métricas de qualidade dos modelos de classificação resultantes. As melhorias superam 10% de aumento na acurácia, de 75% sem nenhum *data augmentation* para valores acima de 85%, incluindo valores de precisão acima de 90%. Os ganhos se mostraram mais acentuados quando quantidades menores de dados rotulados estavam disponíveis, uma vez que os modelos supervisionados sem dados sintéticos são mais propensos a *overfitting* nesses cenários.

Resultados mais detalhados podem ser encontrados na tabela 2. Nela, é possível ver que não só as técnicas de aumento resultam em resultados muito melhores do que a aplicação sem aumento de dados, ela também se aproxima muito da acurácia obtida pelo uso de 100% dos dados disponíveis.

Também é possivel ver que as técnicas de finetuning durante o aumento prejudicaram os resultados de maneira expressiva, em virtude da técnica utilizada e da quantidade de dados disponiveis e, sendo assim, não foi utilizada na etapa final de experimentos.

Outra observação importante é o custo computacional do pre-processamento: apesar das melhoras em qualidade, o tempo para geração das sentenças tem a mesma ordem de grandeza ou é até mesmo maior do que o tempo de treinamento, variando de 30 minutos a várias horas dependendo da quantidade de dados e parametrização. Esse fator pode se tornar um limitante em conjuntos de dados onde o tempo de processamento seja crítico. No entanto, considerando que essa técnica performa proporcionalmente melhor com conjuntos

| **Técnica de Aumento** | **Técnica de classificação** | **Porcentagem de dados rotulados** | **Taxa de aumento de dados** | **top k** | **Acurácia** |
|---|---|---|---|---|---|
| Sem Augmentation | BERT com finetuning | 100% | 0% | 0 | 92% |
| Sem Augmentation | BERT com finetuning | 5% | 0% | 0 | 76% |
| Substituição c/ NPLAug e BERT sem finetuning | BERT com finetuning | 5% | 100% | 50 | 84% |
| Substituição c/ NPLAug e BERT sem finetuning | BERT com finetuning | 5% | 1200% | 5-200 | 87% |
| Substituição c/ NPLAug e BERT com finetuning | BERT com finetuning | 5% | 1200% | 5-200 | 62% |

Tabela 2 – Comparação de resultados preliminares entre modelos sem aumento de dados e com diversas parametrizações do modelo de aumento de dados baseado em substituição de palavras com BERT, com diversos valores de "top k"utilizados para se selecionar a palavra substituida. Em todos eles, apenas 5% dos dados disponíveis tiveram seus rótulos mantidos. Todas as técnicas utilizaram 5 épocas de treinamento com taxa de aprendizado de 0.00002.

menores de dados rotulados, e que a quantidade de dados rotulados é o principal fator que influencia no tempo total, essa limitação pode ser minimizada caso se considere apenas os cenários onde se há mais ganhos.

Finalmente, testes preliminares foram feitos com um conjunto não rotulado de ligações telefônicas automáticas, com o intuito de se classificar campanhas por tipo ou natureza. Uma vez que rótulos não eram conhecidos, especialistas na área foram consultados para agrupar manualmente exemplos, e uma técnica de detecção de tópicos e agrupamento não-supervisionado foi aplicada, chamada BERTopic (GROOTENDORST, 2022). Nessa aplicação, observou-se uma melhora considerável na capacidade de generalização dos modelos de agrupamento treinados: o vocabulário usado deixou de ser o critério principal utilizado no agrupamento dos exemplos, e o modelo passou a corretamente classificar acima de 80% dos exemplos fornecidos por especialista (contra cerca de 50% a 60% do modelo sem aumento de dados). Os experimentos foram omitidos deste documento por terem sido realizados utilizando, parcialmente, informações proprietárias.

Outro ponto importante a ser observado é que, apesar da melhoria quantificável na qualidade dos modelos treinados, os textos de entrada nem sempre se assemelham a textos humanos e, em muitos casos, podem ser identificados como textos artificiais. Isso indica que, apesar da técnica proposta ter se mostrado útil no treinamento de modelos de classificação, seu uso no treinamento de aplicações que se atentam em identificar respostas naturais ou nas quais a consistência das sentenças é importante pode levar a resultados negativos em qualidade.

## 4.1 Cenários de Avaliação

Os experimentos foram divididos em dois cenários principais: um no qual há uma quantidade muito restrita de dados com rótulos (5% do total), e um segundo no qual há uma quantidade maior de rótulos disponíveis (20% do total). O primeiro cenário representa uma situação onde o custo de criar rótulos é proibitivo, ao passo que o segundo representa situações em que rótulos existem, mas não de maneira completa. Ambos os cenários foram gerados sobre o mesmo conjunto de dados.

Em cada um dos cenários, diversas parametrizações foram utilizadas para o aumento de dados, sempre trabalhados sobre o dataset primário:

- O aumento de dados foi feito com diversos fatores de aumento (augmentation factors, 'af'), variando de 1 a 10. Para cada fator de aumento, uma versão aumentada de cada exemplo rotulado foi gerada. O valor de 'af' 0 também foi utilizado como exemplo base - isto é, sem aumento de dados.
- Além disso, o número de palavras consideradas para substituição, chamado 'k' na tabela X, também foi variado. Esse número representa quão diversas ou desconectadas

ao contexto as palavras substituídas/incluídas são, e representa uma posição máxima num ranking: um valor 'k' de 10 indica que uma das 10 palavras mais prováveis foi utilizada, enquanto um 'k' de 150 indica que o espaço possível para seleção de palavras foi maior. Esse valor foi variado de 10 a 150.

- Outros fatores, como a quantidade miníma e máxima de palavras que podem ser modificadas, foram explorados em testes preliminares e tiveram seus valores fixados.

Em relação ao modelo de predição, os seguintes parâmetros foram explorados:

- O Fator de aprendizado (learning rate, "lr") foi brevemente explorado, mas valores tradicionais (na faixa de 1e-5) foram utilizados.
- O uso de finetuning não-supervisionado com os dados não-rotulados foi testado. Quando esse parâmetro é utilizado, todos os dados disponíveis para treinamento (com ou sem rótulos) são apresentados para o treinamento do modelo de classificação. Quando desabilitado, apenas os dados rotulados para treinamento são apresentados.
- O número de épocas de treinamento foi testado em dois cenários: com 5 e 20 épocas. O primeiro cenário representa a relevância dessa técnica em cenários onde o treinamento constante é necessário, ou onde há limitações de recursos para treinamento. Essa abordagem também foi feita para observar a ocorrência de overfitting entre os casos com mais épocas.

Utilizando os parâmetros acima descritos, testes foram realizados com diversas combinações, sobre o dataset 'Covid Twitter' descrito anteriormente. Os resultados completos estão disponíveis no Apêndice A, enquanto uma análise dos mesmos se encontra na seção seguinte.

| percentage | epochs | augmentation_factor | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 5 | 0 | 17.0 | 0.719588 | 0.065331 | 0.560 | 0.67200 | 0.7450 | 0.77600 | 0.794 |
| 5 | 5 | 1 | 3.0 | 0.710000 | 0.125491 | 0.568 | 0.66200 | 0.7560 | 0.78100 | 0.806 |
| 5 | 5 | 2 | 3.0 | 0.826333 | 0.008505 | 0.820 | 0.82150 | 0.8230 | 0.82950 | 0.836 |
| 5 | 5 | 5 | 4.0 | 0.841250 | 0.007676 | 0.833 | 0.83675 | 0.8405 | 0.84500 | 0.851 |
| 5 | 5 | 10 | 4.0 | 0.847500 | 0.004655 | 0.841 | 0.84625 | 0.8485 | 0.84975 | 0.852 |
| 5 | 20 | 0 | 33.0 | 0.835030 | 0.023407 | 0.786 | 0.82000 | 0.8350 | 0.85100 | 0.881 |
| 5 | 20 | 1 | 10.0 | 0.846100 | 0.016086 | 0.820 | 0.83525 | 0.8470 | 0.85675 | 0.871 |
| 5 | 20 | 2 | 16.0 | 0.780000 | 0.135505 | 0.508 | 0.82225 | 0.8415 | 0.85000 | 0.860 |
| 5 | 20 | 5 | 10.0 | 0.858900 | 0.005567 | 0.850 | 0.85600 | 0.8575 | 0.86150 | 0.869 |
| 5 | 20 | 10 | 20.0 | 0.852200 | 0.010866 | 0.832 | 0.84600 | 0.8480 | 0.85650 | 0.875 |
| 20 | 5 | 0 | 1.0 | 0.902000 | - | 0.902 | 0.90200 | 0.9020 | 0.90200 | 0.902 |
| 20 | 5 | 2 | 4.0 | 0.910750 | 0.006551 | 0.901 | 0.91000 | 0.9135 | 0.91425 | 0.915 |
| 20 | 5 | 5 | 4.0 | 0.913500 | 0.010847 | 0.901 | 0.90625 | 0.9145 | 0.92175 | 0.924 |
| 20 | 20 | 0 | 1.0 | 0.907000 | - | 0.907 | 0.90700 | 0.9070 | 0.90700 | 0.907 |
| 20 | 20 | 2 | 4.0 | 0.910750 | 0.008221 | 0.900 | 0.90825 | 0.9115 | 0.91400 | 0.920 |
| 20 | 20 | 5 | 6.0 | 0.910500 | 0.007342 | 0.901 | 0.90700 | 0.9105 | 0.91175 | 0.923 |

Tabela 3 – Resultados dos experimentos agregados por porcentagem dos dados usada, número de épocas de treinamento e fator de aumento (um fator de 0 representa o caso base, sem aumento de dados)

## 4.2 Discussão dos Resultados

Uma comparação dos resultados agregados poder ser vista na tabela 3, contendo as principais métricas para agregações sobre os parâmetros anteriormente descritos.

### 4.2.1 Acurácia em relação ao fator de aumento

Os gráficos 2, 3, 4 e 5 mostram a comparação da acurácia obtida por fator de aumento.

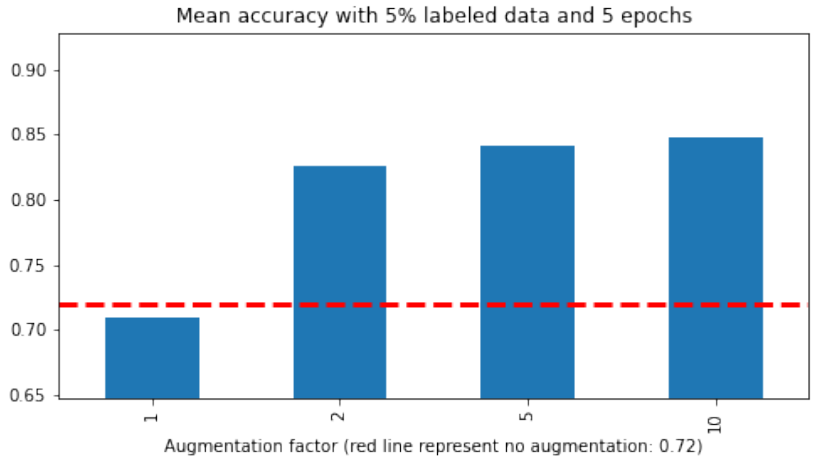


Figura 2 – Acurácia média com 5% de dados rotulados e 5 épocas de treinamento vs fator de aumento (linha vermelha representa a acurácia base, sem aumento)

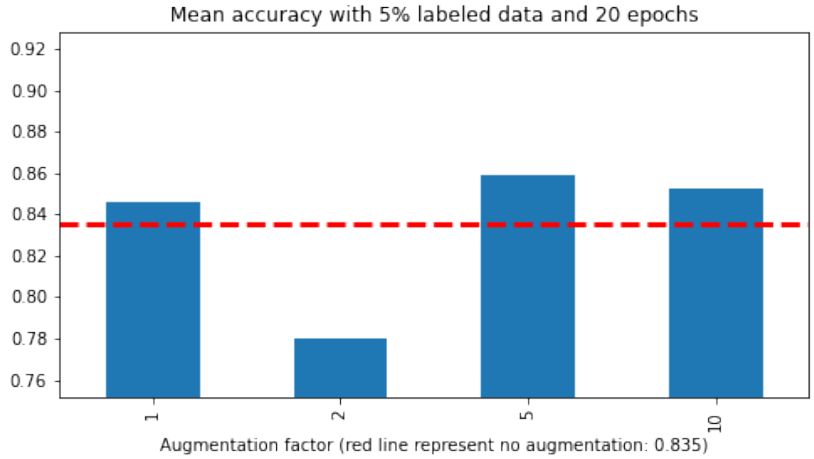


Figura 3 – Acurácia média com 5% de dados rotulados e 20 épocas de treinamento vs fator de aumento (linha vermelha representa a acurácia base, sem aumento)

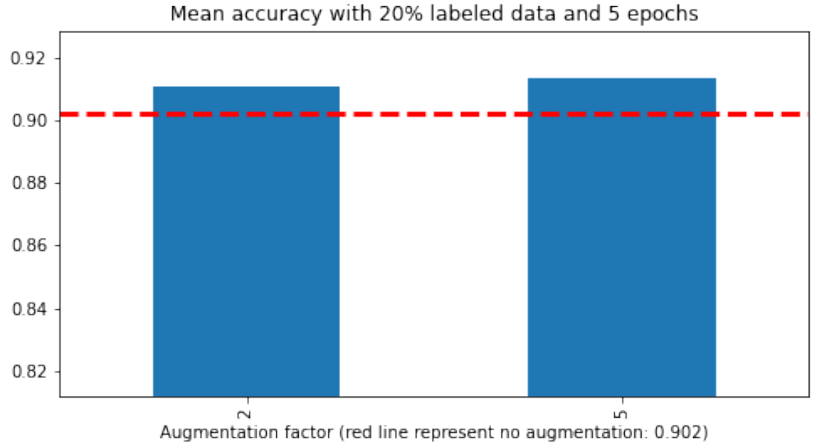


Figura 4 – Acurácia média com 20% de dados rotulados e 5 épocas de treinamento vs fator de aumento (linha vermelha representa a acurácia base, sem aumento)

### 4.2.2 Acurácia em relação ao parâmetro 'k'

Os gráficos 6, 7, 8 e 9 mostram a comparação da acurácia obtida por top 'k' palavras consideradas para substituição.

### 4.2.3 Acurácia por tipo de parametrização

Os gráficos 10, 11, 12 e 13, bem como a tabela 4 mostram a comparação da acurácia sobre o valor médio obtido com aumento de dados, o valor médio objetivo entre as melhores parametrizações de aumento de dados, e os melhores valores incluindo todos os parâmetros.

Baseado nos gráficos e tabelas acima apresentados, é possível ver que o aumento de textos utilizando a técnica proposta levou a uma melhoria de acurácia, especialmente quando há mais limitações ao conjunto de treinamento (como menos dados disponíveis),

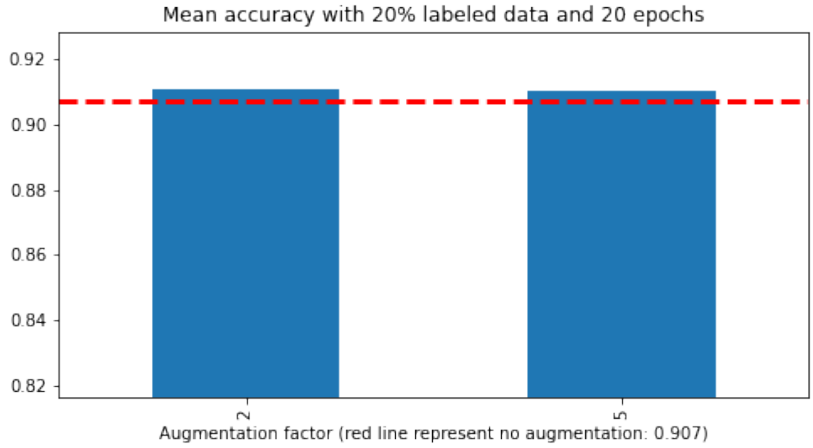


Figura 5 – Acurácia média com 20% de dados rotulados e 20 épocas de treinamento vs fator de aumento (linha vermelha representa a acurácia base, sem aumento)

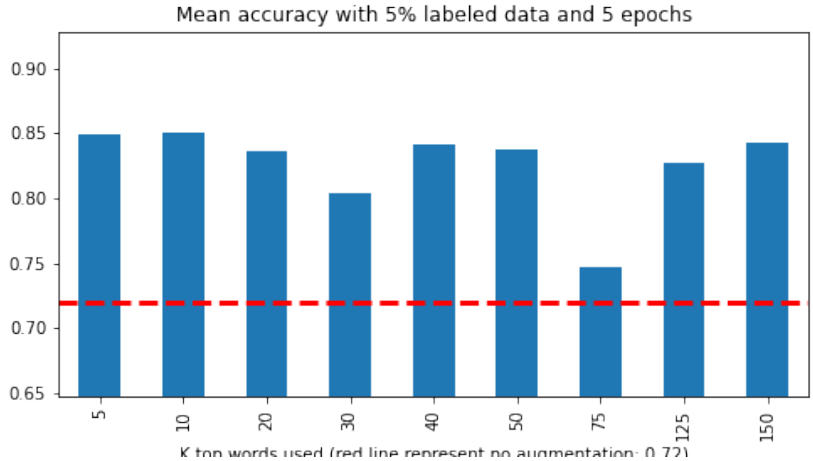


Figura 6 – Acurácia média com 5% de dados rotulados e 5 épocas de treinamento vs 'k' palavras consideradas para substituição (linha vermelha representa a acurácia base, sem aumento)

ao tempo de processamento (com menos épocas) e quando há mais variações disponíveis (maior fator de aumento). Este comportamento é esperado, visto que os padrões com maior viés encontrados no conjunto não-aumentado são mais rapidamente aprendidos, levando a piores resultados que são gradualmente corrigidos com maior tempo de treinamento ou exposição a mais dados. Os modelos treinados com dados aumentados, por sua vez, foram expostos a dados mais variados e com maior riqueza, que não trazem padrões enviesados com a mesma frequência e dominância que o conjunto original.

No cenário 1, com apenas 5% dos dados rotulados, o uso de aumento de dados levou a uma melhoria significativa, especialmente quando o número de épocas é pequeno e o fator de aumento é mais alto. Essa melhoria ultrapassou os 12% de ganho bruto (18% relativo) nas melhores parametrizações, um aumento significativo.

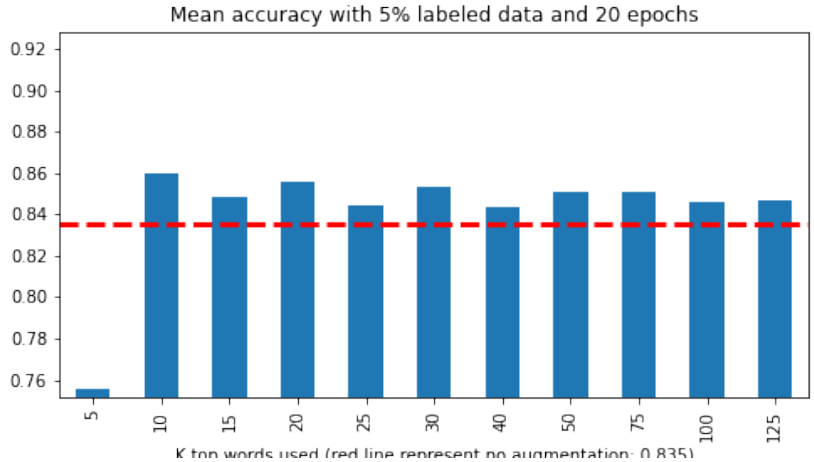


Figura 7 – Acurácia média com 5% de dados rotulados e 20 épocas de treinamento vs 'k' palavras consideradas para substituição (linha vermelha representa a acurácia base, sem aumento)

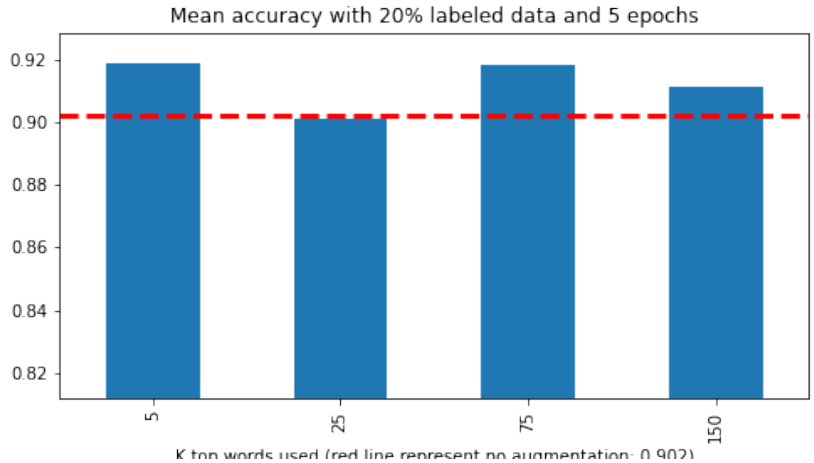


Figura 8 – Acurácia média com 20% de dados rotulados e 5 épocas de treinamento vs 'k' palavras consideradas para substituição (linha vermelha representa a acurácia base, sem aumento)

Ainda no cenário 1, é possível ver que o ganho é reduzido quando mais épocas são consideradas, mas as melhores parametrizações de aumento consistentemente superam as versões não-aumentadas. Também foi possível observar que não há relação clara entre o parâmetro k e a melhoria da acurácia.

No cenário 2, onde 20% dos dados foram utilizados com rótulos, o ganho na utilização de aumento de dados é consideravelmente reduzido, aproximando-se de apenas 2% nas melhores parametrizações. Ainda assim, esse ganho se mostrou estável em diversos experimentos, com o aumento de dados consistentemente superando a técnica sem aumento.

Finalmente, também foi observado que, para experimentos com 20 épocas, os experimentos sem aumento de dados consistentemente atingiram acurácias superiores

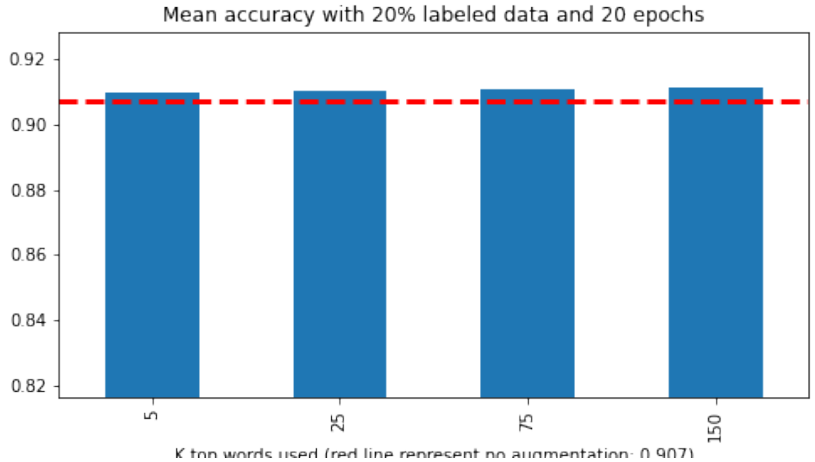


Figura 9 – Acurácia média com 20% de dados rotulados e 20 épocas de treinamento vs 'k' palavras consideradas para substituição (linha vermelha representa a acurácia base, sem aumento)

| percentage | epochs | Baseline | Mean Accuracy | Mean Accuracy (top aug. factor) | Mean Accuracy (best parameters) |
|---|---|---|---|---|---|
| 5 | 5 | 0.719588 | 0.811714 | 0.84750 | 0.852 |
| 5 | 20 | 0.835030 | 0.831679 | 0.85890 | 0.875 |
| 20 | 5 | 0.902000 | 0.912125 | 0.91350 | 0.924 |
| 20 | 20 | 0.907000 | 0.910600 | 0.91075 | 0.923 |

Tabela 4 – Acurácia dos experimentos agregados por porcentagem dos dados usada e número de épocas de treinamento, considerando todos os parâmetros, apenas os melhores fatores de aumento, e apenas as melhores parametrizações.

| | method | 5p_5e | 5p_20e | 20p_5e | 20p_20e | 5p_5e_rank | 5p_20e_rank | 20p_5e_rank | 20p_20e_rank | average_rank |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Caso Base (sem aumento) | 0.719588 | 0.835030 | 0.902000 | 0.90700 | 1.0 | 2.0 | 1.0 | 1.0 | 0.625 |
| **1** | Acurácia Média | 0.811714 | 0.831679 | 0.912125 | 0.91060 | 2.0 | 1.0 | 2.0 | 2.0 | 0.875 |
| **2** | Acurácia Média (melhores fatores de aumento) | 0.847500 | 0.858900 | 0.913500 | 0.91075 | 3.0 | 3.0 | 3.0 | 3.0 | 1.500 |
| **3** | Acurácia Média (melhores parâmetros) | 0.852000 | 0.869000 | 0.924000 | 0.92300 | 4.0 | 4.0 | 4.0 | 4.0 | 2.000 |

Tabela 5 – Comparação de rankeamento entre as técnicas aplicadas

a 95% (muitas vezes atingindo valores próximos de 100%) no conjunto de treinamento, enquanto os experimentos com aumento de dados apresentaram acurácias inferiores - mais próximas ou até inferiores ao do conjunto de validação.

Esses valores são justificados pela presença de palavras e frases mais diversas e que, dada a natureza aleatória da seleção de substituições de palavras, pode ter levado a pares (entrada, rótulo) incongruentes ou conflitantes, mas que ainda assim permitiram uma exposição mais rica do modelo de treinamento ao vocabulário completo do corpo não-rotulado.

### 4.2.4 Análise comparativa

É possível comparar técnicas distintas baseado na colocação comparativa ('rank') das mesmas em diversos cenários. Considerando os dois cenários principais (com 5% e 20%

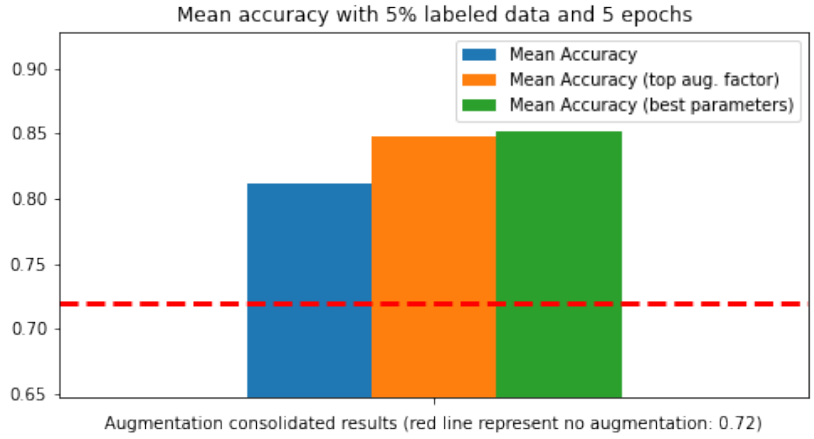


Figura 10 – Acurácia média com 5% de dados rotulados e 5 épocas de treinamento para diversas parametrizações (linha vermelha representa a acurácia base, sem aumento)

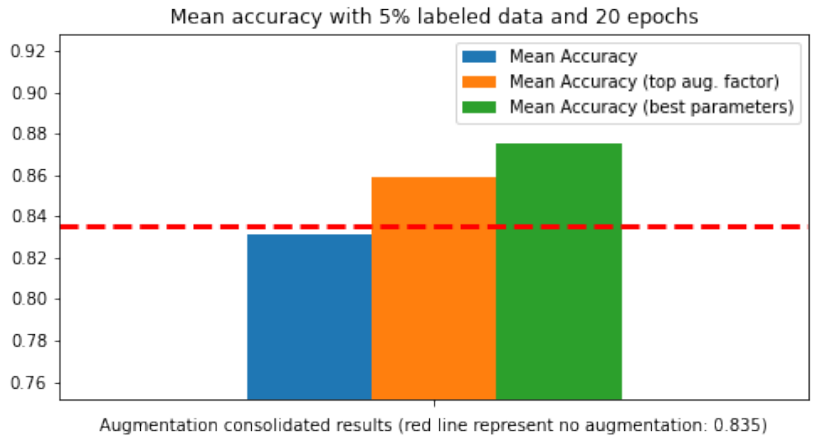


Figura 11 – Acurácia média com 5% de dados rotulados e 20 épocas de treinamento para diversas parametrizações (linha vermelha representa a acurácia base, sem aumento)

dos dados rotulados) e suas variações com 5 e 20 épocas de treinamento, bem como as agregações de métodos analisadas na tabela 4, é possível obter as posições elencadas na tabela 5.

Utilizando-se o teste de Friedman, cujo resultado pode ser visto da figura , não é possível determinar que uma abordagem é superior a outra. Apesar disso, verificou-se que as técnicas com aumento, especialmente quando devidamente parametrizadas, frequentemente mostraram resultados superiores.

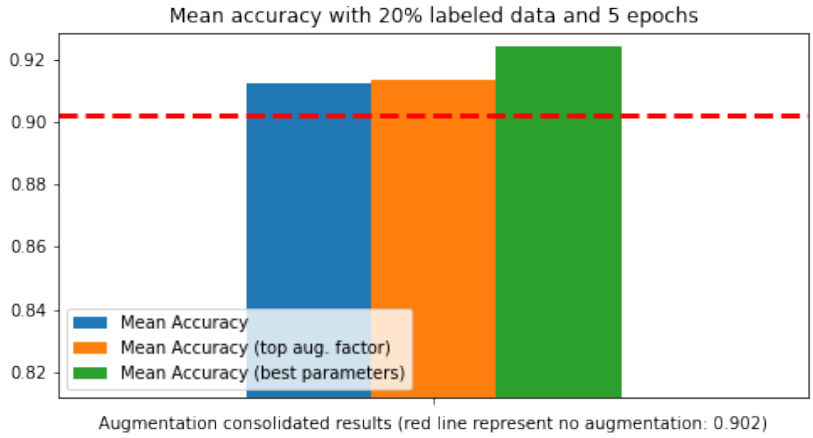


Figura 12 – Acurácia média com 20% de dados rotulados e 5 épocas de treinamento para diversas parametrizações (linha vermelha representa a acurácia base, sem aumento)

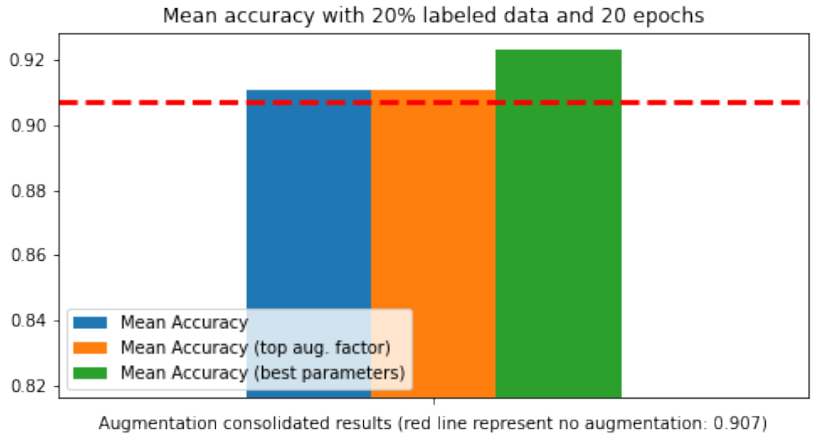


Figura 13 – Acurácia média com 20% de dados rotulados e 20 épocas de treinamento para diversas parametrizações (linha vermelha representa a acurácia base, sem aumento)

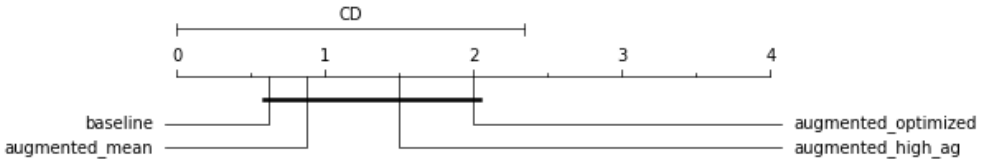


Figura 14 – Resultados do teste de Friedman de diferença crítica)

# 5 CONSIDERAÇÕES FINAIS E CONCLUSÕES

Neste trabalho, diversas abordagens de aumento de dados para conjuntos semi-rotulados de textos foram estudadas e verificou-se que há um ganho real na acurácia de modelos que utilizam dados aumentados sinteticamente. Esse ganho é mais significativo quando uma parcela muito pequena dos dados possui rótulos, e quando há limitações de recursos para o treinamento de modelos de classificação.

## 5.1 Limitações da técnica proposta

Apesar da melhoria nos resultados, a mesma foi bastante pequena em cenários onde mais dados estavam disponíveis para o treinamento, fazendo com que a aplicação dessa técnica possa não ser a melhor solução quando a quantidade de dados rotulados é suficiente para o treinamento de modelos convencionais.

Os testes estatísticos aplicados mostraram que não há ganho significativo na abordagem proposta quando todos os cenários são aplicados, indicando que, por um lado, a técnica proposta tem um escopo de atuação mais restrito do que o anteriormente proposto; e, por outro, que há espaço para novos estudos e abordagens que possam permitir resultados similarmente positivos em outros cenários.

Outro fator que deve ser considerado é o custo computacional da geração dos dados: apesar de não depender do treinamento de modelos de redes neurais artificiais (entre as técnicas empregadas), o aumento de dados exige a execução de modelos pré-treinados que, por sua vez, também requerem poder computacional.

Finalmente, o domínio no qual essas técnicas são aplicadas tende a influenciar a qualidade das expressões aumentadas. Por exemplo, assuntos excessivamente técnicos podem requerer o treinamento de modelos específicos de geração de textos para aumento de dados. Experimentos preliminares com o treinamento de modelos de geração foram feitos, mas não levaram a bons resultados nas técnicas abordadas.

## 5.2 Trabalhos futuros

As técnicas propostas podem ser testadas em outros domínios de texto (como conversas telefônicas ou documentação de código) e aplicações (como classificação de assuntos ou de natureza do texto) e verificar em que circunstancias o aumento de textos é eficaz.

Também é importante considerar que outras técnicas tanto de classificação de textos como de aumento de textos podem ser exploradas, como a utilização de modelos baseados em GPT e outros modelos da família do modelo BERT, uma vez que vários desses

modelos levam a resultados diferentes e utilizaram conjuntos de treinamento distintos. Outras abordagens de aumento (como a adição de sentenças inteiras ou a alteração de estilo de textos para forçar certos rótulos) também podem ser consideradas, especialmente em conjunto com a exploração de outros domínios de dados específicos, umas vez que cada problema de texto possui peculiaridades únicas que podem levar a técnicas distintas com melhores resultados.

Em especifico, o finetuning e pré-treinamento das técnicas de geração e substituição de textos sintéticos podem permitir a obtenção de resultados ainda mais robustos, bem como expandir a aplicação desse tipo de técnica a domínios não cobertos pelo conjunto de treinamento utilizado pelos modelos pré-treinados.

## 5.3 Conclusões finais

A área de aumento de dados pode ser uma aliada valiosa para a solução de inúmeros problemas que contam com restrições na obtenção de rótulos e fazer com que domínios previamente inviáveis de serem resolvidos por inteligência artificial tenham soluções automáticas, seguras e eficientes.

Os ganhos em acurácia apresentados mostram uma melhoria consistente, ainda que pequena, quando os resultados são comparados com a não utilização de técnicas de aumento de dados. Se somada a outras técnicas, essa abordagem tem o potencial de reduzir custos com a criação manual de dados, assim como a redução dos perigos relacionados à rotulação manual, como viés humano e perda de privacidade.

Os experimentos também mostraram que há espaço para melhoria e desenvolvimento de técnicas mais robustas e precisas para o aumento de dados, que permitam enriquecer conjuntos de treinamento e trazer construções sintáticas mais similares ao universo real (não rotulado) de exemplos. Esse desafio envolve a exploração de outros modelos textuais para substituição e síntese de palavras e sentenças, mas pode incluir também melhorias no finetuning de modelos de classificação com conjuntos não-rotulados.

Em resumo, apesar do escopo limitado do trabalho apresentado, a aplicação de técnicas semi-supervisionadas com o intuito de maximizar o uso de informações disponíveis de maneira eficiente é uma área promissora, em especial quando aliada às técnicas de síntese artificial de textos. Com os recentes avanços nessa segunda frente, é possível que novos modelos textuais levem a alternativas sintéticas menos custosas, mais precisas e, eventualmente, robustas o suficiente para permitir que problemas de difícil resolução por limitações na rotulação possam ser solucionados computacionalmente.

# REFERÊNCIAS


AGGARWAL, C. C. **Data mining: the textbook**. [*S.l.: s.n.*]: Springer, 2015.

AGGARWAL, C. C. **Machine learning for text**. [*S.l.: s.n.*]: Springer, 2018.

AGGARWAL, C. C.; ZHAI, C. **Mining text data**. [*S.l.: s.n.*]: Springer Science & Business Media, 2012.

BROWN, T. B. *et al.* Language models are few-shot learners. **arXiv preprint arXiv:2005.14165**, 2020.

CHAPELLE, O.; SCHOLKOPF, B.; ZIEN, A. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. **IEEE Transactions on Neural Networks**, IEEE, v. 20, n. 3, p. 542–542, 2009.

DEVLIN, J. *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

ENGELEN, J. E. v.; HOOS, H. H. Semi-supervised co-ensembling for automl. *In*: SPRINGER. **International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning**. [*S.l.: s.n.*], 2020. p. 229–250.

FENG, S. Y. *et al.* A survey of data augmentation approaches for nlp. **arXiv preprint arXiv:2105.03075**, 2021.

GEORGE, K. S.; JOSEPH, S. Text classification by augmenting bag of words (bow) representation with co-occurrence feature. **IOSR J. Comput. Eng**, v. 16, n. 1, p. 34–38, 2014.

GROOTENDORST, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. **arXiv preprint arXiv:2203.05794**, 2022.

KWON, I.-J. *et al.* Coldexpand: Semi-supervised graph learning in cold start. 2020.

LIU, B. *et al.* When machine learning meets privacy: A survey and outlook. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 54, n. 2, p. 1–36, 2021.

LIU, P. *et al.* A survey of text data augmentation. *In*: IEEE. **2020 International Conference on Computer Communication and Network Security (CCNS)**. [*S.l.: s.n.*], 2020. p. 191–195.

MA, E. **nlpaug: Data augmentation for NLP**. 2019.

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.: s.n.*], 2020. p. 385–399.

MIKOLOV, T. *et al.* Distributed representations of words and phrases and their compositionality. *In*: **Advances in neural information processing systems**. [*S.l.: s.n.*], 2013. p. 3111–3119.

NING, X. *et al.* A review of research on co-training. **Concurrency and computation: practice and experience**, Wiley Online Library, p. e6276, 2021.

NISHI, K. *et al.* Augmentation strategies for learning with noisy labels. *In*: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [*S.l.: s.n.*], 2021. p. 8022–8031.

OTTER, D. W.; MEDINA, J. R.; KALITA, J. K. A survey of the usages of deep learning for natural language processing. **IEEE Transactions on Neural Networks and Learning Systems**, IEEE, v. 32, n. 2, p. 604–624, 2020.

PETERS, M. *et al.* **Deep contextualized word representations. NAACL-HLT**. [*S.l.: s.n.*]: Association for Computational Linguistics2227–2237, 2018.

RADFORD, A. *et al.* Improving language understanding by generative pre-training. 2018.

RADFORD, A. *et al.* Language models are unsupervised multitask learners. **OpenAI blog**, v. 1, n. 8, p. 9, 2019.

SHORTEN, C.; KHOSHGOFTAAR, T. M.; FURHT, B. Text data augmentation for deep learning. **Journal of big Data**, Springer, v. 8, n. 1, p. 1–34, 2021.

SIDOROV, G. *et al.* Syntactic n-grams as machine learning features for natural language processing. **Expert Systems with Applications**, Elsevier, v. 41, n. 3, p. 853–860, 2014.

SUNDERMEYER, M.; SCHLÜTER, R.; NEY, H. Lstm neural networks for language modeling. *In*: **Thirteenth annual conference of the international speech communication association**. [*S.l.: s.n.*], 2012.

TORREY, L.; SHAVLIK, J. Transfer learning. *In*: **Handbook of research on machine learning applications and trends: algorithms, methods, and techniques**. [*S.l.: s.n.*]: IGI global, 2010. p. 242–264.

VAJJALA, S. *et al.* **Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems**. [*S.l.: s.n.*]: O'Reilly Media, 2020.

VASWANI, A. *et al.* Attention is all you need. *In*: **Advances in neural information processing systems**. [*S.l.: s.n.*], 2017. p. 5998–6008.

YANG, Z. *et al.* Xlnet: Generalized autoregressive pretraining for language understanding. **Advances in neural information processing systems**, v. 32, 2019.

YAROWSKY, D. Unsupervised word sense disambiguation rivaling supervised methods. *In*: **33rd annual meeting of the association for computational linguistics**. [*S.l.: s.n.*], 1995. p. 189–196.

ZAKI, M. J.; JR, W. M. **Data Mining and Machine Learning: Fundamental Concepts and Algorithms**. [*S.l.: s.n.*]: Cambridge University Press, 2020.

ZHU, X.; GOLDBERG, A. B. Introduction to semi-supervised learning. **Synthesis lectures on artificial intelligence and machine learning**, Morgan & Claypool Publishers, v. 3, n. 1, p. 1–130, 2009.

# APÊNDICES

# APÊNDICE A – EXPERIMENTOS COMPLETOS

| k | Fator de Aumento | finetuning | percentage | p__min | p__max | p__p | epochs | Taxa de Aprendizado | tp | tn | fp | fn | accuracy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 2 | False | 5 | 5 | 10 | 9 | 20 | 2e-05 | 0 | 508 | 492 | 0 | 0.508 |
| 5 | 2 | False | 5 | 5 | 10 | 9 | 20 | 2e-05 | 0 | 508 | 492 | 0 | 0.508 |
| 5 | 2 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 0 | 508 | 492 | 0 | 0.508 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 5 | 1e-05 | 179 | 381 | 313 | 127 | 0.56 |
| 75 | 1 | True | 5 | 5 | 10 | 9 | 5 | 2e-05 | 69 | 499 | 423 | 9 | 0.568 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 5 | 1e-05 | 177 | 458 | 315 | 50 | 0.635 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 5 | 1e-05 | 238 | 429 | 254 | 79 | 0.667 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 5 | 1e-05 | 201 | 467 | 291 | 41 | 0.668 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 5 | 1e-05 | 255 | 417 | 237 | 91 | 0.672 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 5 | 1e-05 | 270 | 415 | 222 | 93 | 0.685 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 5 | 1e-05 | 366 | 334 | 126 | 174 | 0.7 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 5 | 1e-05 | 301 | 408 | 191 | 100 | 0.709 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 5 | 2e-05 | 290 | 455 | 202 | 53 | 0.745 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 5 | 2e-05 | 310 | 439 | 182 | 69 | 0.749 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 5 | 2e-05 | 309 | 444 | 183 | 64 | 0.753 |
| 30 | 1 | True | 5 | 5 | 10 | 9 | 5 | 2e-05 | 355 | 401 | 137 | 107 | 0.756 |
| 5 | 0 | False | 5 | 5 | 10 | 9 | 5 | 2e-05 | 386 | 382 | 106 | 126 | 0.768 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 5 | 2e-05 | 347 | 429 | 145 | 79 | 0.776 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 5 | 2e-05 | 340 | 436 | 152 | 72 | 0.776 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 5 | 2e-05 | 379 | 404 | 113 | 104 | 0.783 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 1e-05 | 390 | 396 | 102 | 112 | 0.786 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 5 | 2e-05 | 350 | 443 | 142 | 65 | 0.793 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 5 | 2e-05 | 355 | 439 | 137 | 69 | 0.794 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 1e-05 | 380 | 423 | 112 | 85 | 0.803 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 1e-05 | 416 | 387 | 76 | 121 | 0.803 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 1e-05 | 403 | 401 | 89 | 107 | 0.804 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 2e-05 | 348 | 456 | 144 | 52 | 0.804 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 1e-05 | 382 | 422 | 110 | 86 | 0.804 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 1e-05 | 382 | 423 | 110 | 85 | 0.805 |
| 125 | 1 | True | 5 | 5 | 10 | 9 | 5 | 2e-05 | 425 | 381 | 67 | 127 | 0.806 |
| 5 | 2 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 408 | 403 | 84 | 105 | 0.811 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 1e-05 | 387 | 432 | 105 | 76 | 0.819 |
| 75 | 2 | True | 5 | 5 | 10 | 9 | 5 | 2e-05 | 401 | 419 | 91 | 89 | 0.82 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 2e-05 | 392 | 428 | 100 | 80 | 0.82 |
| 40 | 1 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 390 | 430 | 102 | 78 | 0.82 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 1e-05 | 382 | 440 | 110 | 68 | 0.822 |
| 30 | 2 | True | 5 | 5 | 10 | 9 | 5 | 2e-05 | 366 | 457 | 126 | 51 | 0.823 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 2e-05 | 385 | 438 | 107 | 70 | 0.823 |
| 50 | 2 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 411 | 415 | 81 | 93 | 0.826 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 1e-05 | 368 | 459 | 124 | 49 | 0.827 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 1e-05 | 384 | 444 | 108 | 64 | 0.828 |
| 75 | 1 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 403 | 427 | 89 | 81 | 0.83 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 2e-05 | 411 | 420 | 81 | 88 | 0.831 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 2e-05 | 410 | 421 | 82 | 87 | 0.831 |
| 25 | 10 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 392 | 440 | 100 | 68 | 0.832 |
| 30 | 5 | True | 5 | 5 | 10 | 9 | 5 | 2e-05 | 435 | 398 | 57 | 110 | 0.833 |
| 25 | 2 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 411 | 422 | 81 | 86 | 0.833 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 2e-05 | 405 | 430 | 87 | 78 | 0.835 |

**Tabela 6 continuação da página anterior**

| k | Fator de Aumento | finetuning | percentage | p__min | p__max | p__p | epochs | Taxa de Aprendizado | tp | tn | fp | fn | accuracy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 1e-05 | 385 | 450 | 107 | 58 | 0.835 |
| 20 | 1 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 395 | 440 | 97 | 68 | 0.835 |
| 20 | 2 | True | 5 | 5 | 10 | 9 | 5 | 2e-05 | 443 | 393 | 49 | 115 | 0.836 |
| 100 | 1 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 398 | 438 | 94 | 70 | 0.836 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 1e-05 | 409 | 429 | 83 | 79 | 0.838 |
| 50 | 5 | True | 5 | 5 | 10 | 9 | 5 | 2e-05 | 436 | 402 | 56 | 106 | 0.838 |
| 40 | 2 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 395 | 444 | 97 | 64 | 0.839 |
| 25 | 10 | True | 5 | 5 | 10 | 9 | 20 | 1e-05 | 421 | 419 | 71 | 89 | 0.84 |
| 75 | 2 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 397 | 444 | 95 | 64 | 0.841 |
| 15 | 1 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 434 | 407 | 58 | 101 | 0.841 |
| 40 | 10 | True | 5 | 5 | 10 | 9 | 5 | 2e-05 | 408 | 433 | 84 | 75 | 0.841 |
| 15 | 2 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 432 | 410 | 60 | 98 | 0.842 |
| 100 | 2 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 414 | 428 | 78 | 80 | 0.842 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 2e-05 | 432 | 410 | 60 | 98 | 0.842 |
| 150 | 5 | True | 5 | 5 | 10 | 9 | 5 | 2e-05 | 418 | 425 | 74 | 83 | 0.843 |
| 50 | 10 | True | 5 | 5 | 10 | 9 | 20 | 1e-05 | 415 | 429 | 77 | 79 | 0.844 |
| 5 | 10 | True | 5 | 5 | 10 | 9 | 20 | 1e-05 | 413 | 432 | 79 | 76 | 0.845 |
| 75 | 10 | True | 5 | 5 | 10 | 9 | 20 | 1e-05 | 421 | 425 | 71 | 83 | 0.846 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 1e-05 | 418 | 428 | 74 | 80 | 0.846 |
| 15 | 10 | True | 5 | 5 | 10 | 9 | 20 | 1e-05 | 417 | 429 | 75 | 79 | 0.846 |
| 40 | 10 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 403 | 443 | 89 | 65 | 0.846 |
| 125 | 10 | True | 5 | 5 | 10 | 9 | 20 | 1e-05 | 417 | 430 | 75 | 78 | 0.847 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 2e-05 | 431 | 416 | 61 | 92 | 0.847 |
| 5 | 10 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 423 | 424 | 69 | 84 | 0.847 |
| 30 | 10 | True | 5 | 5 | 10 | 9 | 20 | 1e-05 | 423 | 424 | 69 | 84 | 0.847 |
| 30 | 2 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 421 | 426 | 71 | 82 | 0.847 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 1e-05 | 406 | 441 | 86 | 67 | 0.847 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 2e-05 | 414 | 433 | 78 | 75 | 0.847 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 2e-05 | 433 | 415 | 59 | 93 | 0.848 |
| 125 | 10 | True | 5 | 5 | 10 | 9 | 5 | 2e-05 | 440 | 408 | 52 | 100 | 0.848 |
| 5 | 10 | True | 5 | 5 | 10 | 9 | 5 | 2e-05 | 414 | 435 | 78 | 73 | 0.849 |
| 10 | 10 | True | 5 | 5 | 10 | 9 | 20 | 1e-05 | 417 | 432 | 75 | 76 | 0.849 |
| 5 | 2 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 432 | 417 | 60 | 91 | 0.849 |
| 100 | 5 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 425 | 425 | 67 | 83 | 0.85 |
| 10 | 5 | True | 5 | 5 | 10 | 9 | 5 | 2e-05 | 412 | 439 | 80 | 69 | 0.851 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 2e-05 | 417 | 434 | 75 | 74 | 0.851 |
| 20 | 10 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 443 | 409 | 49 | 99 | 0.852 |
| 75 | 10 | True | 5 | 5 | 10 | 9 | 5 | 2e-05 | 410 | 442 | 82 | 66 | 0.852 |
| 5 | 2 | False | 5 | 5 | 10 | 9 | 20 | 2e-05 | 417 | 436 | 75 | 72 | 0.853 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 1e-05 | 426 | 427 | 66 | 81 | 0.853 |
| 30 | 1 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 419 | 434 | 73 | 74 | 0.853 |
| 5 | 2 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 411 | 443 | 81 | 65 | 0.854 |
| 40 | 10 | True | 5 | 5 | 10 | 9 | 20 | 1e-05 | 432 | 422 | 60 | 86 | 0.854 |
| 20 | 5 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 455 | 400 | 37 | 108 | 0.855 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 2e-05 | 425 | 430 | 67 | 78 | 0.855 |
| 15 | 10 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 432 | 424 | 60 | 84 | 0.856 |
| 30 | 5 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 420 | 436 | 72 | 72 | 0.856 |
| 100 | 10 | True | 5 | 5 | 10 | 9 | 20 | 1e-05 | 420 | 436 | 72 | 72 | 0.856 |
| 25 | 1 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 420 | 436 | 72 | 72 | 0.856 |
| 50 | 5 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 421 | 435 | 71 | 73 | 0.856 |
| 50 | 1 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 439 | 418 | 53 | 90 | 0.857 |
| 15 | 5 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 419 | 438 | 73 | 70 | 0.857 |

**Tabela 6 continuação da página anterior**

| k | Fator de Aumento | finetuning | percentage | p__min | p__max | p__p | epochs | Taxa de Aprendizado | tp | tn | fp | fn | accuracy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 2e-05 | 424 | 433 | 68 | 75 | 0.857 |
| 10 | 10 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 413 | 445 | 79 | 63 | 0.858 |
| 40 | 5 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 415 | 443 | 77 | 65 | 0.858 |
| 10 | 0 | False | 5 | 5 | 10 | 9 | 20 | 1e-05 | 431 | 427 | 61 | 81 | 0.858 |
| 10 | 2 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 423 | 436 | 69 | 72 | 0.859 |
| 20 | 2 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 429 | 431 | 63 | 77 | 0.86 |
| 25 | 5 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 439 | 421 | 53 | 87 | 0.86 |
| 5 | 1 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 431 | 431 | 61 | 77 | 0.862 |
| 10 | 5 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 427 | 435 | 65 | 73 | 0.862 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 1e-05 | 422 | 441 | 70 | 67 | 0.863 |
| 30 | 10 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 436 | 428 | 56 | 80 | 0.864 |
| 5 | 5 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 427 | 439 | 65 | 69 | 0.866 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 2e-05 | 425 | 443 | 67 | 65 | 0.868 |
| 75 | 10 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 420 | 449 | 72 | 59 | 0.869 |
| 75 | 5 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 427 | 442 | 65 | 66 | 0.869 |
| 50 | 10 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 439 | 432 | 53 | 76 | 0.871 |
| 10 | 1 | True | 5 | 5 | 10 | 9 | 20 | 2e-05 | 436 | 435 | 56 | 73 | 0.871 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 2e-05 | 438 | 437 | 54 | 71 | 0.875 |
| 20 | 10 | True | 5 | 5 | 10 | 9 | 20 | 1e-05 | 436 | 439 | 56 | 69 | 0.875 |
| 0 | 0 | False | 5 | 0 | 0 | 0 | 20 | 2e-05 | 445 | 436 | 47 | 72 | 0.881 |
| 5 | 2 | True | 20 | 5 | 10 | 9 | 20 | 1e-05 | 452 | 448 | 40 | 60 | 0.9 |
| 25 | 5 | True | 20 | 5 | 10 | 9 | 20 | 1e-05 | 451 | 450 | 41 | 58 | 0.901 |
| 25 | 5 | True | 20 | 5 | 10 | 9 | 5 | 2e-05 | 433 | 468 | 59 | 40 | 0.901 |
| 25 | 2 | True | 20 | 5 | 10 | 9 | 5 | 2e-05 | 447 | 454 | 45 | 54 | 0.901 |
| 5 | 0 | True | 20 | 5 | 10 | 9 | 5 | 2e-05 | 453 | 449 | 39 | 59 | 0.902 |
| 5 | 5 | True | 20 | 5 | 10 | 9 | 20 | 2e-05 | 441 | 465 | 51 | 43 | 0.906 |
| 10 | 0 | True | 20 | 5 | 10 | 9 | 20 | 2e-05 | 456 | 451 | 36 | 57 | 0.907 |
| 150 | 5 | True | 20 | 5 | 10 | 9 | 5 | 2e-05 | 457 | 451 | 35 | 57 | 0.908 |
| 75 | 5 | True | 20 | 5 | 10 | 9 | 20 | 1e-05 | 449 | 461 | 43 | 47 | 0.91 |
| 150 | 2 | True | 20 | 5 | 10 | 9 | 20 | 1e-05 | 449 | 462 | 43 | 46 | 0.911 |
| 5 | 5 | True | 20 | 5 | 10 | 9 | 20 | 1e-05 | 453 | 458 | 39 | 50 | 0.911 |
| 75 | 2 | True | 20 | 5 | 10 | 9 | 20 | 1e-05 | 458 | 454 | 34 | 54 | 0.912 |
| 150 | 5 | True | 20 | 5 | 10 | 9 | 20 | 1e-05 | 447 | 465 | 45 | 43 | 0.912 |
| 5 | 2 | True | 20 | 5 | 10 | 9 | 5 | 2e-05 | 457 | 456 | 35 | 52 | 0.913 |
| 150 | 2 | True | 20 | 5 | 10 | 9 | 5 | 2e-05 | 448 | 466 | 44 | 42 | 0.914 |
| 75 | 2 | True | 20 | 5 | 10 | 9 | 5 | 2e-05 | 456 | 459 | 36 | 49 | 0.915 |
| 25 | 2 | True | 20 | 5 | 10 | 9 | 20 | 1e-05 | 459 | 461 | 33 | 47 | 0.92 |
| 75 | 5 | True | 20 | 5 | 10 | 9 | 5 | 2e-05 | 460 | 461 | 32 | 47 | 0.921 |
| 5 | 5 | False | 20 | 5 | 10 | 9 | 20 | 2e-05 | 448 | 475 | 44 | 33 | 0.923 |
| 5 | 5 | True | 20 | 5 | 10 | 9 | 5 | 2e-05 | 464 | 460 | 28 | 48 | 0.924 |

Tabela 6 – Experimentos completos