UNIVERSIDADE FEDERAL DO MARANHÃO

UNIVERSIDADE FEDERAL DO PIAUÍ

Doutorado em Ciência da Computação Associação UFMA/UFPI

Hélcio de Abreu Soares

Detecção de Correlações Espúrias com Inteligência Artificial Explicável

Orientador: Prof. Dr. Rodrigo de Melo Souza Veras

Coorientador: Prof. Dr. Raimundo Santos Moura

Teresina - PI

Setembro, 2025

Hélcio de Abreu Soares

Detecção de Correlações Espúrias com Inteligência Artificial Explicável

TESE DE DOUTORADO

A Tese apresentada como requisito parcial para obtenção do título de Doutor em Ciência da Computação, ao Doutorado em Ciência da Computação, Associação UFMA/UFPI.

Orientador: Prof. Dr. Rodrigo de Melo Souza Veras
Coorientador: Prof. Dr. Raimundo Santos Moura

Teresina - PI
Setembro, 2025

FICHA CATALOGRÁFICA
Universidade Federal do Piauí
Biblioteca Comunitária Jornalista Carlos Castello Branco
Serviço de Processos Técnicos

S676d Soares, Hélcio de Abreu.
Detecção de correlações espúrias com inteligência artificial explicável / Hélcio de Abreu Soares. – 2025.
107 f.

Tese (Doutorado) – Universidade Federal do Piauí / Universidade Federal do Maranhão, Doutorado em Ciência da Computação, Teresina, 2025.
“Orientador: Prof. Dr. Rodrigo de Melo Souza Veras.”
“Coorientador: Prof. Dr. Raimundo Santos Moura.”

1. Inteligência artificial. 2. Classificação binária. 3. Padrões espúrios. 4. PLN. 5. XAI. I. Veras, Rodrigo de Melo Souza. II. Moura, Raimundo Santos. III. Título.

CDD 006.32

Bibliotecário: Gésio dos Santos Barros - CRB3/1469

Hélcio de Abreu Soares

Detecção de Correlações Espúrias com Inteligência Artificial Explicável

A presente Tese de Doutorado foi avaliada por banca examinadora composta pelos seguintes membros:

Prof. Dr. Ajalmar Rego da Rocha Neto
Instituto Federal do Ceará

Prof. Dr. Anselmo Cardoso de Paiva
Universidade Federal do Maranhão

Prof. Dr. Gustavo Paiva Guedes e Silva
Centro Federal de Educação Tecnológica Celso Suckow da Fonseca

Prof. Dr. Vinícius Ponte Machado
Universidade Federal do Piauí

Certificamos que esta é a versão original e final da Tese de Doutorado que foi julgada adequada para a obtenção do título de Doutor em Ciência da Computação

gov.br Documento assinado digitalmente
RODRIGO DE MELO SOUZA VERAS
Data: 24/10/2025 14:25:30-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Rodrigo de Melo Souza Veras
Orientador

gov.br Documento assinado digitalmente
ANSELMO CARDOSO DE PAIVA
Data: 31/10/2025 16:11:48-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Anselmo Cardoso de Paiva
Coordenador

Teresina - PI, 26 Setembro 2025

Aos meus pais, Luzia Maria de Abreu Soares e Pedro Oliveira Soares (in memoriam) por suas palavras de incentivo, amor incondicional e constante encorajamento. A conclusão dessa etapa é também um reflexo da dedicação e valores que vocês me transmitiram.

Agradecimentos

É com imensa gratidão que expresso meus sinceros agradecimentos às pessoas e instituições que estiveram ao meu lado, proporcionando apoio, orientação e inspiração ao longo deste percurso desafiador.

Primeiramente, expresso minha profunda gratidão a Deus, cuja graça e orientação estiveram presentes a cada passo deste caminho. Sua sabedoria infinita e amor incondicional foram minha fonte de força nos momentos de desafio e dúvida.

Aos meus queridos filhos, Lucas Rodrigues Soares e Amanda Rodrigues Soares por entender minha ausência em muitos momentos. Cada um de vocês é uma fonte constante de inspiração e motivação, e meu desejo é que este trabalho possa contribuir de alguma forma para um mundo melhor para vocês e para as gerações futuras.

Ao Tribunal de Contas do Piauí, expresso minha gratidão por ter me proporcionado a oportunidade de realizar este estudo. O apoio e o ambiente propício para a pesquisa foram fundamentais para o desenvolvimento deste trabalho.

Da mesma forma, gostaria de expressar minha profunda gratidão ao meu orientador, Prof. Dr. Rodrigo de Melo Souza Veras, cuja orientação experiente, paciência, sabedoria e incentivo constantes foram fundamentais para a realização deste trabalho.

Gostaria também de estender meus agradecimentos aos membros da banca examinadora, por dedicarem seu tempo, expertise e análises criteriosas a este trabalho. Suas observações, críticas construtivas e sugestões serão essenciais para aprimorar a qualidade desta pesquisa.

A todos aqueles que, de alguma forma, contribuíram para esta jornada, seja com conselhos, apoio emocional ou incentivo, meu mais sincero obrigado. Esta conquista não seria possível sem vocês.

“A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê.”

(Arthur Schopenhauer)

Resumo

Apesar dos avanços em Inteligência Artificial (IA), modelos de *Machine Learning* e *Deep Learning* ainda carecem de transparência e explicabilidade, sendo tratados como “caixas-pretas”. Este trabalho aborda o problema das correlações espúrias — associações entre padrões e classes sem relação causal — que, em tarefas de classificação binária em Processamento de Linguagem Natural (PLN), comprometem a precisão, a imparcialidade e a generalização dos modelos. Propomos um método que combina técnicas de Inteligência Artificial Explicável (XAI) e aprendizado não supervisionado para identificar e graduar padrões espúrios. Utilizando o algoritmo *K-means*, os padrões são agrupados e analisados pela distância aos centroides, sob a hipótese de que distâncias maiores indicam maior grau de espuriedade. A abordagem considera a influência desses padrões sobre explicadores e sua associação com erros de previsão. A metodologia é aplicada a dados de licitações e contratos do Tribunal de Contas do Estado do Piauí (TCE-PI), usando modelos baseados em *Support Vector Machine* (SVM), *Logistic Regression* (LR) com representações textuais TF-IDF e *Word Embeddings*, e o modelo BERTimbau, como codificador e classificador com *embeddings* contextuais dinâmicos. Aplicamos também o método ao IMDB para avaliar generalização e compará-lo com métodos de referências. Os resultados confirmam a hipótese e mostram consistência entre modelos e bases. As principais contribuições incluem: (i) método agnóstico a modelos e explicadores; (ii) detecção automática de padrões espúrios; (iii) uma métrica de espuriedade baseada na distância ao centroide; e (iv) organização lógica e interpretável dos padrões, ampliando a compreensão dos modelos e apoiando a mitigação de padrões espúrios.

Palavras-chave: PLN; XAI; Classificação binária; Padrões espúrios.

Abstract

Despite advances in Artificial Intelligence (AI), Machine Learning and Deep Learning models still lack transparency and explainability, often being regarded as “black boxes.” This dissertation addresses the issue of spurious correlations—associations between patterns and classes that lack causal relationships—which, in binary classification tasks in Natural Language Processing (NLP), undermine model accuracy, fairness, and generalization. We propose a method that combines Explainable Artificial Intelligence (XAI) techniques with unsupervised learning to identify and rank spurious patterns. Using the K-means algorithm, patterns are clustered and evaluated based on their distance from centroids under the hypothesis that greater distances indicate higher degrees of spuriousness. The approach accounts for the influence of these patterns on explainers and their association with prediction errors. The methodology is applied to procurement and contract data from the Court of Auditors of the State of Piauí (TCE-PI), using Support Vector Machines (SVM), Logistic Regression with TF-IDF and Word Embedding text representations, and the BERTimbau model, both as encoder and classifier with dynamic contextual embeddings. The method is also applied to the IMDB dataset to evaluate generalization and compare it against reference methods. The results confirm the hypothesis and reveal consistent patterns across models and datasets. The main contributions include: (i) a model- and explainer-agnostic method; (ii) automatic detection of spurious patterns; (iii) a spuriousness metric based on centroid distance; and (iv) logical and interpretable organization of patterns, enhancing model understanding and supporting the mitigation of spurious correlations.

Keywords: NLP; XAI; Binary classification; Spurious patterns.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AEC	*Argument Extraction Corpus*
AMT	*Amazon Mechanical Turk*
AUC	*Area Under the Curve*
BOW	*Bag-of-Words*
BERT	*Bidirectional Encoder Representations from Transformers*
DASED	*Debater Argument Search Engine Dataset*
DRO	*Distributionally Robust Optimization*
EDSF	*Explainability-based Detection of Spurious Features*
EASP	*Environment-Agnostic Sequential Predictor*
ERM	*Empirical Risk Minimization*
IA	*Inteligência Artificial*
IFL	*Invariant Feature Learning*
INVRAT	*Invariant Rationalization*
IRM	*Invariant Risk Minimization*
LR	*Logistic Regression*
MMI	*Maximum Mutual Information*
MLP	*Multilayer Perceptron*
MT-DNN	*Multi-Task Deep Neural Network*
NDCG	*Normalized Discounted Cumulative Gain*

PCA	*Principal Component Analysis*
PLN	*Processamento de Linguagem Natural*
ROC	*Receiver Operating Characteristics*
RNA	*Rede Neural Artificial*
SCAVI	*Source Code Analysis for Vulnerability Identification*
SST-2	*Stanford Sentiment Treebank*
SVD	*Singular Value Decomposition*
SVM	*Support Vector Machines*
SGD	*Stochastic Gradient Descent*
XAI	*Explainable Artificial Intelligence*
SHAP	*SHapley Additive exPlanations*
SSL	*Self-Supervised Learning*
LIME	*Local Interpretable Model-agnostic Explanations*
TF-IDF	*Term Frequency-Inverse Document Frequency*

Sumário

1 Introdução

Apesar dos avanços tecnológicos na utilização da Inteligência Artificial (IA), grande parte dos algoritmos, incluindo Redes Neurais Artificiais (RNA), Máquinas de Vetores de Suporte e Regressão Logística, ainda são considerados “caixas pretas”, devido à falta de transparência, interpretabilidade e explicabilidade. Essa característica gera inseguranças e incertezas, que necessitam saber: como um modelo chega a uma decisão específica; como um conjunto de entradas pode levar à produção de uma saída; por que o modelo errou e se as associações entre entradas e saídas refletem causas reais ou apenas correlações espúrias. Uma correlação espúria refere-se a uma conexão entre duas variáveis que parece ser causal, mas não é. No contexto de IA, correlações espúrias podem levar a previsões imprecisas devido às associações estatísticas nos dados de treinamento que não refletem as verdadeiras relações causais (Yuan et al. 2023).

Correlações espúrias podem ter origem em artefatos de dados, que são características presentes nos dados de treinamento que podem ser erroneamente aprendidas pelo modelo como sendo relevantes para sua tarefa, mas que na verdade não têm relação causal com a classe alvo (Gautam et al. 2023). Esses artefatos podem surgir de um viés de dados, onde, por exemplo, objetos de uma classe específica de imagens frequentemente aparecem com um determinado fundo, levando o modelo a aprender o fundo em vez do objeto (Lapuschkin et al. 2019). Adicionalmente, os dados de treinamento podem ser manipulados por meio da inserção de um gatilho conhecido como “*backdoor*”, que, se presente, sempre resulta na previsão de uma classe específica — efetivamente um atalho para essa classe alvo (Chen et al. 2018). Outro fenômeno, denominado “*Clever Hans*”, ocorre quando um artefato correlacionado com uma classe nos dados de treinamento influencia o modelo a fazer uma previsão correta, mas pelos motivos errados. (Lapuschkin et al. 2019). Outra fonte de correlações espúrias são os confundidores. No contexto de classificação, confundidores são variáveis que podem distorcer a relação

entre a variável independente (atributos observáveis) e a variável dependente (classe) (Keith, Jensen e O'Connor 2020). Esses fatores podem surgir de várias fontes, incluindo a forma como os conjuntos de dados são rotulados ou de suas características inerentes. Na prática, os confundidores podem desviar o modelo da aprendizagem de características verdadeiramente relevantes e robustas, confundindo-as com correlações espúrias (Howell et al. 2021).

Em Processamento de Linguagem Natural (PLN), padrões espúrios são correlações espúrias entre um atributo (ou agrupamentos de atributos) e uma classe, cuja validade não se mantém fora do treinamento (Mu et al. 2022). Se o modelo incorporar padrões espúrios durante o treinamento, ele terá uma alta probabilidade de falha (Schwartz e Stanovsky 2022). Esses padrões podem levar a resultados enganosos, induzindo problemas como viés, falhas na generalização e dificuldades de interpretação (Pezeshkpour et al. 2021). Tais correlações podem resultar em um modelo acertando por razões erradas ou falhando por influências dos padrões espúrios. Wang e Culotta 2020b demonstram que um modelo pode aprender a associar indevidamente o termo "Spielberg" a críticas positivas, não pela qualidade dos filmes, mas porque títulos dirigidos por Spielberg tendem a receber mais avaliações positivas no treinamento. Assim, o classificador pode rotular incorretamente resenhas positivas sem o termo como negativas e vice-versa, ignorando o real sentimento expresso no texto.

Em Processamento de Imagens, (Lapuschkin et al. 2019) analisam o fenômeno "Clever Hans", onde um modelo aprende padrões irrelevantes em vez da tarefa principal. No conjunto de dados LISA, um artefato visual (quadrado amarelo) foi inserido em todas as imagens de sinais de parada. O modelo treinado obteve 100% de acurácia nessas imagens, mas apenas 6,5% nas que não continham o artefato, evidenciando sua dependência do quadrado amarelo em vez das características reais dos sinais.

Esses exemplos ilustram o risco de modelos aprenderem a se basear em características que não têm relação causal com a tarefa desejada. Esses erros e falsos acertos comprometem a confiabilidade, a interpretabilidade e a aplicabilidade prática do modelo em situações reais, onde a precisão e a

justiça das decisões são necessárias. Entretanto, quando padrões espúrios são detectados, compreendidos e adequadamente tratados, há um grande potencial para melhorar a robustez e a confiabilidade dos modelos, assegurando que suas previsões reflitam padrões verdadeiramente relevantes. Portanto, a má compreensão de como um modelo funciona pode levar à baixa confiança e ao risco de implementação de uma solução que faz julgamentos baseados em padrões espúrios. Detectar e compreender esses padrões aumenta a robustez e a confiabilidade no modelo. Assim, é necessário estar ciente dos pontos fortes e fracos do modelo, para que saibamos interpretar seus resultados e, consequentemente, melhorar o seu desempenho.

Técnicas de Inteligência Artificial Explicável (*Explainable Artificial Intelligence - XAI*) facilitam a compreensão e a confiança humana nos resultados gerados por algoritmos de IA (Arrieta et al. 2020). Essas técnicas envolvem processos e métodos que tornam as previsões dos modelos mais compreensíveis, estabelecendo uma relação clara entre os atributos de entrada e as decisões tomadas. Isso aumenta a transparência dos modelos e reduz o problema da “caixa preta”. Além disso, essas características podem auxiliar na detecção de artefatos de dados e confundidores (Wang et al. 2021).

Diante desses desafios, torna-se necessário desenvolver estratégias que não apenas detectem padrões espúrios, mas também permitam avaliá-los de forma quantitativa e interpretável. Com base nessa motivação, definimos os objetivos que norteiam esta pesquisa.

1.1 Objetivos

Este trabalho propõe um método que combina a importância de recursos baseada em XAI com aprendizado não supervisionado. O objetivo é identificar e agrupar potenciais padrões espúrios de maneira interpretável, permitindo uma análise quantitativa de sua relação com as classes de saída em tarefas de classificação binária. Investigamos a conexão entre *outliers* e padrões espúrios, formulando a hipótese de que *outliers* podem representar esses padrões. Embora o conceito de *outlier* seja bem estabelecido, sua associação

com padrões espúrios ainda não foi investigada. Propomos que a distância dos padrões ao centro do *cluster* (centroide), gerados pelo algoritmo *K-means*, sirva como critério para avaliar seu grau de espuriedade, partindo da premissa de que padrões mais distantes podem indicar associações instáveis ou inconsistentes. Essa métrica permite mensurar de forma objetiva a influência de cada padrão nas previsões dos modelos.

A partir dessa hipótese, elaboramos uma heurística que usa métricas estatísticas de interações modelo-dados, fundamentando-se em evidências que demonstram a influência de tais padrões sobre os explicadores (Plumb, Ribeiro e Talwalkar 2021, Chou et al. 2022, Cardozo et al. 2022, Anders et al. 2022, Srivastava 2023, Gautam et al. 2023), sejam artefatos ou confundidores. Além disso, fundamenta-se nos argumentos de padrões espúrios são uma das causas de erros de previsão em modelos (Wang e Culotta 2020b, Izmailov et al. 2022, Schwartz e Stanovsky 2022, Du et al. 2022, Ali et al. 2023, Kumar, Deshpande e Sharma 2023, Chew et al. 2024), e que a dificuldade dos modelos em classificar corretamente objetos deve-se, também, à presença de objetos semelhantes, mas com rótulos diferentes, nos dados de treinamento (Kattakinda e Feizi 2021, Wu et al. 2023, Zhou et al. 2023, Tjandra e Wiens 2023). Diante dessas evidências, desenvolvemos uma abordagem que explora a interação entre explicações globais, padrões frequentes em erros e influência nas previsões, para gerar *clusters* lógicos e interpretáveis através do *K-means*.

Aplicamos a abordagem proposta a dois conjuntos de dados do Tribunal de Contas do Estado do Piauí (TCE-PI), referentes a editais de licitações e a contratos públicos. Utilizamos modelos de *Machine Learning* e *Deep Learning*, explorando representações textuais tanto baseadas em frequência (TF-IDF) quanto contextuais (*Word Embeddings*). Para avaliar a capacidade de generalização, estendemos os testes à base Internet Movie Database (IMDB) e comparamos os resultados com métodos de referência da literatura.

Os resultados indicam que a distância dos padrões aos centroides dos agrupamentos pode ser utilizada como métrica para quantificar o grau de dependência do modelo em relação a esses padrões — sendo que distâncias maiores correspondem a maior dependência. Além disso, os agrupamentos

revelam particularidades distintas associadas aos modelos e às representações textuais utilizadas, destacando o impacto dessas escolhas na aprendizagem de padrões e oferecendo subsídios concretos para a aplicação de estratégias de mitigação de correlações espúrias.

1.2 Objetivos Específicos

Com o intuito de alcançar o objetivo central deste estudo, foram delineados objetivos específicos necessários. A consecução desses objetivos específicos culminará no alcance do objetivo central por esta pesquisa; são eles:

- Desenvolver um método para detectar padrões espúrios em classificações binárias, combinando explicabilidade com aprendizado não supervisionado sobre métricas de importância, frequência e desempenho.
- Propor e validar uma métrica baseada na distância ao centroide para estimar a espuriedade dos padrões, analisando sua relação com os erros de classificação por meio de perturbações e desempenho.
- Testar a abordagem nas bases do TCE-PI e IMDB, comparando com métodos de referência para avaliar sua generalização.
- Comparar modelos e representações textuais, analisando a influência de padrões espúrios e o comportamento em diferentes contextos.

1.3 Contribuição e Produção Científica

Diferente de abordagens anteriores que focam em palavras isoladas, este método detecta padrões compostos por múltiplas palavras. Sua originalidade está na combinação simultânea de três características: (*i*) detecção automatizada de padrões espúrios, sem intervenção manual; (*ii*) uso da distância aos centroides como métrica para o potencial espúrio de padrões; e (*iii*) identificação de padrões compostos. A combinação desses itens orientada à quantificação e interpretabilidade da espuriedade, configura a principal inovação metodológica deste trabalho. As principais contribuições são:

- Um método agnóstico ao modelo para detectar e quantificar padrões espúrios;
- Abordagem automatizada e não supervisionada, sem necessidade de seleção manual de candidatos a padrões espúrios.
- Agrupamento lógico e interpretável de padrões com uso da distância ao centroide como métrica de espuriedade.
- Código do método, instruções de uso e conjuntos de dados reais de objetos de licitações e contratos públicos para classificação binária e análise de correlações espúrias disponíveis em: www.github.com/HelcioSoares-IFPI/SpuriousPatternTracker[1]

Além dessas contribuições, este trabalho resultou em publicações científicas que reforçam sua relevância e aplicabilidade. As produções estão organizadas a seguir, em ordem cronológica:

- **Soares, H.**, Veras, R., Moura, R., Paiva, A. *Using Explainability to Find Spurious Patterns in Textual Datasets*. In: **Intelligent Systems Design and Applications (ISDA)**, Springer, 2023, pp. 424–434.
- **Soares, H.**, Moura, R., Machado, V. P., Paiva, A., Lima, W., Veras, R. *The Detection of Spurious Correlations in Public Bidding and Contract Descriptions Using Explainable Artificial Intelligence and Unsupervised Learning*. **Electronics**, MDPI, vol. 14, n. 7, p. 1251, 2025.

Este documento está organizado da seguinte forma: No Capítulo 2 explora os fundamentos teóricos, abordando conceitos, recursos e ferramentas da pesquisa. No Capítulo 3, discutimos uma revisão da literatura, fornecendo uma perspectiva sobre os principais trabalhos na área de estudo. No Capítulo 4, detalhamos as abordagens do trabalho, a metodologia e a base de dados. O Capítulo 5 detalha os experimentos com as abordagens propostas, acompanhados de análises e discussões. Por fim, o Capítulo 6 apresenta conclusões, limitações e sugestões para futuras pesquisas.

[1] Acessado em 14 de dezembro de 2024.

2 Referencial Teórico

Este capítulo apresenta os conceitos e técnicas que fundamentam o estudo, abordando Processamento de Linguagem Natural (PLN), Inteligência Artificial Explicável (XAI) e Aprendizado Não Supervisionado.

2.1 Processamento de Linguagem Natural (PLN)

O PLN é uma subárea da inteligência artificial voltada à interação entre computadores e linguagem humana. No contexto de classificação, o PLN envolve a utilização de algoritmos e modelos para categorizar automaticamente textos em diferentes classes ou categorias, com base em seu conteúdo (Chopra, Prashar e Sain 2013).

2.1.1 Modelos de Processamento de Linguagem Natural (PLN)

A maioria dos problemas de classificação em PLN pode ser formalmente modelada por meio da equação apresentada por Eisenstein 2018:

$$\widehat{y} = \underset{y \in Y(x)}{\arg\max} \ \Psi(x, y; \theta) \tag{2.1}$$

onde, x representa a entrada e y denota uma saída, pertencente ao conjunto de possíveis respostas $Y(x)$. A função $\Psi(x, y; \theta)$ é um escore que quantifica a adequação da saída y à entrada x, parametrizada por θ, $\widehat{y}$ é, portanto, a saída que maximiza a função de escore, refletindo a hipótese mais provável segundo o modelo parametrizado. A formulação é compatível com uma ampla classe de modelos supervisionados em PLN, incluindo classificadores lineares e redes neurais e, em certos contextos, modelos baseados em *transformers* após etapas de ajuste fino. A escolha da função Ψ e dos parâmetros θ influencia diretamente a capacidade preditiva do modelo. No entanto, a interpretabilidade está mais relacionada à complexidade funcional de Ψ e à arquitetura do modelo, exigindo, frequentemente métodos auxiliares para sua análise.

Durante o treinamento, os modelos aprendem a associar textos a categorias usando exemplos. A eficácia dos modelos é influenciada pela qualidade e quantidade dos dados de treinamento, pela complexidade do modelo e sua capacidade de generalizar para textos não vistos (Goodfellow, Bengio e Courville 2016).

Regressão Logística. Modelos baseados em Regressão Logística (*Logistic Regression* - LRG) são usados em aprendizado de máquina supervisionado para classificação binária devido à sua simplicidade, interpretabilidade e eficácia em lidar com grandes conjuntos de dados. Esse algoritmo procura otimizar um conjunto de parâmetros para maximizar a estimativa da probabilidade de uma saída, usando a técnica de descida de gradiente para ajustar os coeficientes da função logística, que é fundamental para calcular as probabilidades de classificação e fornecer interpretações claras dos resultados (Kleinbaum et al. 2002, Harrell et al. 2001). São considerados de explicabilidade intrínseca, pois seus coeficientes podem ser diretamente interpretados como a influência de cada variável de entrada na probabilidade de uma classe de saída. Regressão Logística tem limitações, incluindo a suposição de independência e a relação linear entre as variáveis preditoras e o *logit* da variável resposta, que pode falhar, especialmente em textos com dependências semânticas e sintáticas complexas (Bountakas, Koutroumpouchos e Xenakis 2021). Selecionamos esse algoritmo por sua explicabilidade intrínseca, que se alinha com nossa abordagem e como uma linha de base para comparar o desempenho de modelos mais complexos.

Máquina de Vetores de Suporte com Gradiente Descendente Estocástico. A Máquina de Vetores de Suporte (*Support Vector Machine* - SVM) otimizada com Gradiente Descendente Estocástico (*Stochastic Gradient Descent* - SGD) é um classificador linear utilizado em aprendizado de máquina (Li e Orabona 2019). Diferente das SVMs tradicionais baseadas em kernel, que exploram mapeamentos não lineares dos dados, a SVM com SGD otimiza diretamente uma função de perda *hinge*, priorizando eficiência computacional em grandes volumes de dados textuais, o que o torna escalável para aplicações em larga escala e mantém o desempenho em tarefas de classificação de texto.

Modelos baseados em SVM treinados com SGD enfrentam desafios relacionados à convergência e à estabilidade dos parâmetros. Como o SGD realiza atualizações incrementais nos pesos do modelo a cada instância ou minibatch, pequenas variações nas amostras podem induzir oscilações nos gradientes, afetando a estabilidade das fronteiras de decisão (Wijnhoven e With 2010). A escolha do SVM nesta pesquisa fundamenta-se em evidências empíricas de trabalhos anteriores. Em particular, Soares et al. 2024 mostraram que esse modelo apresentou desempenho satisfatório na detecção de correlações espúrias em conjuntos de dados com características similares aos utilizados neste estudo.

BERT. O *Bidirectional Encoder Representations from Transformers* (BERT) é um modelo de linguagem desenvolvido pelo *Google* em 2018, que utiliza uma abordagem bidirecional para processar texto, permitindo que o contexto de cada palavra seja analisado tanto à esquerda quanto à direita simultaneamente. Esta abordagem representa um avanço significativo em relação aos modelos unidirecionais anteriores, ao incorporar mecanismos de atenção bidirecional baseados em *Transformer Encoder*, permitindo uma compreensão mais profunda do contexto linguístico (Devlin et al. 2018). O treinamento do BERT ocorre em duas fases principais: primeiro, no pré-treinamento, onde o modelo é treinado em grandes corpora de texto não rotulado usando duas tarefas — o *Masked Language Model* (MLM), em que palavras são ocultadas e o modelo aprende a prever com base no contexto, o *Next Sentence Prediction* (NSP), que permite ao modelo prever a continuidade lógica entre sentenças. Em seguida, na fase de ajuste fino, o modelo é adaptado a tarefas específicas de PLN, utilizando conjuntos de dados rotulados menores para aprimorar seu desempenho em tarefas como classificação de texto, resposta a perguntas e reconhecimento de entidades nomeadas. A estratégia de aprendizado por transferência aproveita o pré-treinamento para alcançar um desempenho elevado, mesmo com poucos dados rotulados na etapa de ajuste fino.

Neste estudo, o BERT é utilizado em duas funções: como extrator de *Embeddings* para os modelos lineares e como classificador binário. Ao gerar representações densas e ricas em contexto, o modelo contribui para que os classificadores lineares identifiquem padrões relevantes com maior eficácia. Quando empregado como classificador, permite uma análise de sua sensibilidade a padrões espúrios em comparação com os modelos lineares. Adicionalmente, sua compatibilidade com a técnica de explicabilidade LIME facilita a interpretação das decisões do modelo, aspecto fundamental para os objetivos desta pesquisa.

2.1.2 Representações Textuais

Modelos de PLN necessitam que textos sejam representados de forma compreensível para máquinas. Para isso, diversas representações textuais foram desenvolvidas. Embora haja outras formas de representações textuais, abordaremos apenas aquelas utilizadas neste trabalho, destacando suas relações com os modelos utilizados.

Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF é uma melhoria sobre *Bag-of-Words* (BOW) (Shao et al. 2018), que não apenas conta a frequência de termos (TF) mas também ajusta essas contagens ponderando cada termo pela frequência inversa do documento (IDF). O objetivo é reduzir a importância de palavras comuns e aumentar a importância de palavras raras que podem ser mais discriminatórias em um conjunto de documentos. A fórmula do TF-IDF é dada por:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t) \tag{2.2}$$

onde $\text{TF}(t, d)$ é a frequência do termo t no documento d, e $\text{IDF}(t) = \log\left(\frac{N}{n_t}\right)$.

TF-IDF é útil em modelos que se beneficiam de representações esparsas e discriminativas, como SVM, LRG e Naive Bayes. No entanto, como BoW, TF-IDF não captura relações semânticas e contextuais entre palavras, limitando sua eficácia em modelos de deep learning mais complexos como transformers.

Word Embeddings. São representações densas de palavras em espaços vetoriais contínuos, em que palavras semanticamente similares se encontram próximas. Word2Vec (Mikolov et al. 2013) e Doc2Vec (Le e Mikolov 2014), antecessores dos *embeddings* modernos, introduziram a capacidade de representar a semântica em vetores, abrindo caminho para técnicas mais recentes, entre elas *GloVe* (Pennington, Socher e Manning 2014) e *FastText* (Bojanowski et al. 2017). *GloVe* utiliza estatísticas globais de coocorrência de palavras em um corpus para gerar *embeddings*, enquanto *FastText* incorpora subpalavras, possibilitando um melhor manejo de palavras raras e da morfologia. Os *Word embeddings* são essenciais para modelos baseados em *transformers*, como BERT (Devlin et al. 2018) e GPT-2 (Radford et al. 2019), que utilizam essas representações vetoriais como ponto de partida para capturar relações contextuais complexas em texto. Esses modelos refinam os *embeddings* durante o treinamento para se alinhar às suas tarefas específicas.

Neste estudo, diferenciamos dois tipos de representações. Os modelos lineares (SVM e LRG) utilizaram *Word Embeddings* fixos, doravante denominados **WE**, gerados com o BERTimbau por meio da estrutura Sentence-BERT (`SBERT`), que transforma cada sentença em um vetor denso. Como esses classificadores não refinam os vetores durante o treinamento, os WE permanecem inalterados. Já o modelo BERTimbau, utilizado diretamente como classificador, emprega *embeddings* contextuais, ajustados dinamicamente a cada instância e refinados ao longo do processo de aprendizado, permitindo capturar relações semânticas mais complexas.

2.2 Inteligência Artificial Explicável (XAI)

A Inteligência Artificial Explicável (*eXplainable Artificial Intelligence - XAI*) visa tornar as decisões da IA mais transparentes e compreensíveis. A XAI inclui métodos intrínsecos, integrados ao design do modelo para garantir transparência e métodos *post-hoc*, aplicados após o treinamento para explicar decisões dos modelos. Essas técnicas são classificadas em várias categorias, cada uma abordando diferentes aspectos da explicabilidade. A seguir, uma taxonomia proposta em (Ali et al. 2023).

Baseada no Escopo da Explicação. A análise de importância de características revela a influência das entradas nas saídas do modelo, classificando-se em local ou global. Explicadores Locais fornecem explicações para previsões individuais; *Local Interpretable Model-agnostic Explanations* (LIME), por exemplo, cria modelos interpretáveis em torno de instâncias específicas (Ribeiro, Singh e Guestrin 2016). Em contraste, Explicadores Globais oferecem uma visão geral do comportamento do modelo no conjunto de dados e o *SHapley Additive exPlanations* (SHAP), um exemplo, utiliza valores de *Shapley* da teoria dos jogos para identificar a importância de cada característica, sendo aplicável tanto local quanto globalmente (Ali et al. 2023).

Baseada na Integração com o Modelo. De maneira geral, modelos mais complexos são difíceis de explicar, enquanto os mais simples, embora possam ser interpretáveis, sacrificam a precisão (Breiman 2001, Sengupta et al. 2023, Gaurav e Tiwari 2023). Os *explicadores intrínsecos* integram a explicabilidade ao design do modelo, usando modelos lineares, árvores de decisão e sistemas baseados em regras (Izza et al. 2023), por exemplo. Esses modelos, embora interpretáveis, perdem em precisão (Adadi e Berrada 2018). Já os *explicadores pós-hoc* aplicam técnicas após o treinamento de modelos complexos para fornecer explicações. O *Gradient-weighted Class Activation Mapping* (Grad-CAM), é um exemplo de explicador pós-hoc que usa gradientes para criar mapas de calor, destacando regiões importantes em imagens para modelos de convolução (Wang et al. 2020).

Baseada na Dependência do Modelo. As estratégias de interpretabilidade dividem-se em independentes ou específicas do modelo (Adadi e Berrada 2018). *Explicadores Agnósticos ao Modelo* utilizam técnicas que não dependem de um modelo específico, fornecem explicações pós-hoc local e/ou global (Zafar e Khan 2021). Exemplos incluem o *Individual Conditional Expectation* (ICE) e LIME, que gera explicações criando modelos locais (Goldstein et al. 2015). Já os *Explicadores Específicos ao Modelo*, como o *Layer-wise Relevance Propagation* (LRP), que é exclusivo para redes neurais e analisa a contribuição de cada neurônio de acordo com suas camadas (Hu et al. 2021).

2.3 Aprendizado não supervisionado

O aprendizado não supervisionado analisa dados sem rótulos para identificar padrões e correlações de forma autônoma (Barlow 1989). Diferente do aprendizado supervisionado, trabalha apenas com dados de entrada, sendo útil para descobrir conhecimento e estruturas ocultas (Hastie et al. 2009). Técnicas comuns em aprendizado não supervisionado incluem métodos de agrupamento (*clustering*), entre eles *K-means* e *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), além de abordagens de redução de dimensionalidade, incluindo *Principal Component Analysis* (PCA). Esses algoritmos permitem que modelos identifiquem padrões de entrada com base nas propriedades estatísticas dos dados, sem a necessidade de saídas-alvo explícitas (Ghahramani 2003).

2.3.1 K-means

O *K-means* é um algoritmo de *clustering* amplamente utilizado em aprendizado não supervisionado. Seu objetivo é particionar n observações em k *clusters* de modo que cada observação esteja no *cluster* cujo centroide (média) é mais próximo. Este processo iterativo busca minimizar a variância intra-*cluster*, representada pela soma dos quadrados das distâncias entre os pontos e seus respectivos centroides. Frequentemente, o *K-means* é usado em análise exploratória para identificar grupos homogêneos em conjuntos de dados (Mussabayev et al. 2023). A formulação matemática do objetivo do *K-means* é expressa pela minimização da função objetivo

$$J = \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{2.3}$$

onde μ_i representa o centroide do *cluster* S_i, e x é um ponto de dados dentro de S_i (MacQueen et al. 1967). O algoritmo inicia escolhendo aleatoriamente k pontos de dados como centros iniciais, depois, cada observação é atribuída ao *cluster* cujo centroide está mais próximo. Os centroides são então recalculados como a média de todas as observações atribuídas a cada *cluster*. Este processo é repetido até que os centroides não mudem significativamente entre as

iterações ou até que um número máximo de iterações seja atingido. Um dos desafios é determinar, antecipadamente, o número apropriado de *clusters*, k. Sem uma estimativa precisa, pode-se sub ou superestimar a quantidade ideal de *clusters*, afetando a eficácia do agrupamento.

O *método do cotovelo* é uma técnica heurística utilizada para determinar o número ideal de *clusters*. Consiste em plotar a quantidade de *clusters* no eixo x e a Soma dos Quadrados das Distâncias (SSD) no eixo y. Inicialmente, a SSD diminui rapidamente conforme os *clusters* aumentam, mas atinge um ponto em que a redução se torna insignificante, formando um "cotovelo" na curva. Esse ponto indica o número adequado (Bholowalia e Kumar 2014, MacQueen et al. 1967). A abordagem visual facilita a escolha desse valor, equilibrando complexidade e precisão. Na literatura, é amplamente utilizada como critério preliminar em análises com K-means (Syakur et al. 2018).

2.3.2 Análise de Componentes Principais

A Análise de Componentes Principais (PCA) é um método estatístico que transforma variáveis possivelmente correlacionadas em componentes principais linearmente descorrelacionadas, capturando a variação presente nos dados (Bro e Smilde 2014). A PCA é usada para explorar, visualizar e reduzir a dimensionalidade dos dados, mantendo a maior parte da variação original, identificando padrões e simplificando a complexidade dos dados. O PCA transforma o conjunto original de variáveis em um novo conjunto menor de variáveis (componentes principais), que são combinações lineares das variáveis originais. Esses componentes principais retêm as características essenciais dos dados, capturando a maior variância possível com um número reduzido de componentes (Kambhatla e Leen 1997). A fórmula da Análise de Componentes Principais (PCA) começa com o cálculo da matriz de covariância S a partir da matriz de dados X de dimensão $n \times p$, onde n é o número de observações e p é o número de variáveis. A matriz de covariância é dada por:

$$S = \frac{1}{n-1} X^T X \tag{2.4}$$

onde X^T é a transposta de X, e S é uma matriz $p \times p$ representando as covariâncias entre as variáveis. Em seguida, realiza-se a decomposição de autovalores da matriz S, calculando os autovalores λ e autovetores v pela equação $Sv = \lambda v$. Esta etapa identifica as direções de maior variância nos dados.

Em resumo, este capítulo consolidou os fundamentos teóricos necessários para nossa pesquisa, abordando modelos de classificação, representações textuais, técnicas de explicabilidade e métodos não supervisionados. Esses conceitos estruturam a proposta metodológica e fornecem a base analítica para compreender e quantificar padrões espúrios. A partir dessa fundamentação, o próximo capítulo apresenta os trabalhos relacionados, situando esta pesquisa no estado da arte e destacando as lacunas que motivam a proposta desenvolvida.

3 Trabalhos Relacionados

3.1 Introdução

Neste capítulo apresentamos os principais trabalhos relacionados com o objetivo de compará-los à proposta desenvolvida, destacando semelhanças, diferenças metodológicas e limitações. A seleção dos estudos foi realizada por meio de uma pesquisa sistemática da literatura, utilizando as bases *Scopus*, *Springer Nature* e *IEEE Xplore*. A busca foi conduzida com a chave ("deep learning"OR "machine learning") AND "spurious correlations"AND ("nlp"OR "natural language processing"), considerando apenas artigos publicados entre 2020 e 2024.

Inicialmente, foram identificados 142 artigos. Após a remoção de duplicatas, 125 estudos únicos foram avaliados com base nos seguintes critérios: (i) tratar da detecção de correlações espúrias em dados textuais; (ii) incluir estudos em imagens apenas se influenciaram o método desta Tese; (iii) apresentar uso de explicadores ou estratégias de análise de padrões espúrios; (iv) conter proposta metodológica ou experimento, excluindo *surveys*, revisões e textos teóricos sem validação; e (v) apresentar relevância temática para o tema.

Os trabalhos selecionados foram organizados em quatro grupos: (i) análise e geração de contrafactuais, (ii) perturbação de dados, (iii) explicabilidade (XAI) e (iv) outras técnicas baseadas em representação causal ou otimização robusta. A classificação adotada baseia-se nas abordagens metodológicas predominantes em cada estudo. As técnicas de geração de contrafactuais e perturbação de dados são amplamente utilizadas na literatura para identificar correlações espúrias, pois permitem avaliar diretamente a sensibilidade do modelo a alterações controladas nas entradas. Já os trabalhos baseados em XAI foram agrupados separadamente por estarem diretamente relacionados à hipótese desta Tese, segundo a qual os explicadores podem revelar padrões espúrios considerados relevantes pelos modelos. Por fim, foram incluídas em um

grupo distinto as propostas que utilizam outras estratégias, como representação causal ou métodos de otimização robusta, que não se enquadram nas categorias anteriores.

3.2 Análise e Geração de Contrafactuais

Esta seção reúne estudos que utilizam contrafactuais — instâncias artificiais geradas a partir de modificações controladas em exemplos reais, preservando o conteúdo principal mas alterando aspectos específicos (Chou et al. 2022) — que são inseridas na base de dados original.

O trabalho de Wang e Culotta 2020a propõe aumentar a robustez de classificadores de texto a correlações espúrias via geração automática de contrafactuais. Primeiro, treina-se um classificador linear com regressão logística sobre representações bag-of-words, no qual cada termo recebe um coeficiente proporcional à sua contribuição para a predição. Selecionam-se os termos com maiores coeficientes absolutos, considerados candidatos a espúrios. Para cada termo, aplica-se a técnica *closest opposite matching*, que busca sentenças semanticamente semelhantes com rótulo oposto e sem o termo, usando *embeddings* BERT. Se a similaridade exceder um limiar (ex.: 0.95), o termo é considerado causal. Com auxílio de dicionários, encontram-se antônimos desses termos, gerando contrafactuais ao substituir o termo pelo antônimo e inverter o rótulo. O classificador é então retreinado com os dados originais e os contrafactuais, o que melhora a robustez frente a mudanças na distribuição dos dados. Os autores relatam até 5% de melhora em bases como IMDB e *Amazon Kindle*, especialmente com contrafactuais manuais.

Semelhante à nossa proposta, o trabalho ora examinado utiliza termos considerados importantes para cada classe como candidatos a espúrios e analisa sentenças semelhantes com rótulos opostos. Este último, porém, tem um propósito diferente: enquanto considera todas as sentenças semelhantes que contêm termos candidatos a espúrios, nossa proposta foca apenas nas sentenças semelhantes às que o modelo errou e que compartilham esses termos. Isso significa que nossa análise é mais direcionada aos erros reais

do classificador com base em evidências concretas de falha. Além disso, apresentamos um método de detecção automática de padrões espúrios, sem necessidade de geração de contrafactuais ou intervenção manual. Ao contrário do trabalho ora analisado, que trabalha com alterações de rótulo e requerem intervenções específicas para cada padrão causal identificado, além de explorar apenas palavras individuais.

O trabalho de Wu et al. 2022 propõe um método de geração e filtragem de dados para mitigar correlações espúrias em tarefas de inferência textual (*Natural Language Inference – NLI*), que classifica a relação entre uma premissa e uma hipótese como implicação, contradição ou neutra. Os autores utilizam bases de dados com vieses conhecidos, como conjuntos com exemplos com manipulações sintáticas e testes adversariais. O método parte da geração de dados por meio do modelo GPT-2, que é ajustado para criar exemplos consistentes com os rótulos das tarefas. Para evitar que o gerador reproduza padrões espúrios presentes nos dados originais, os autores propõe a técnica de *unlikelihood training*, que penaliza a geração de exemplos inconsistentes. Em seguida, introduzem o conceito de *z-statistics*, uma medida estatística utilizada para detectar atributos cuja frequência difere significativamente entre as classes. A partir disso, o algoritmo *z-filtering* remove amostras que contenham atributos com altos valores de z associadas a uma determinada classe, sob a suposição de que esses atributos não deveriam influenciar a classificação. O processo é usado para criar versões dos conjuntos de dados originais com menor presença de padrões espúrios. Nos experimentos, modelos BERT-base treinados com os dados gerados e filtrados apresentaram melhor desempenho em conjuntos específicos voltados à detecção de vieses, método superou estratégias anteriores, com ganhos de até 13.3% em acurácia. Os resultados também se mantêm em modelos maiores, como RoBERTa e ALBERT, indicando a robustez da abordagem.

A proposta assemelha-se a nossa Tese por também identificar padrões espúrios com base em seu impacto sobre o desempenho dos modelos. Contudo, há diferenças significativas: (i) o método atua na geração e curadoria dos dados de entrada, enquanto nossa proposta atua após o treinamento do modelo,

identificando padrões espúrios a partir de explicações locais, frequência e análise de erros; (ii) requer predefinição manual dos atributos considerados independentes da tarefa, os quais são usados para calcular os *z-statistics*, nossa abordagem não exige intervenção manual nem conhecimento prévio sobre os atributos; e (iii) nossa abordagem permite detectar padrões compostos e específicos do domínio, enquanto a abordagem apresentada no artigo opera com atributos mais simples.

Yadav et al. 2022 propõem uma abordagem baseada na *Tsetlin Machine (TM)*, um modelo que aprende regras lógicas do tipo `AND`, utilizando tanto termos afirmativos quanto suas negações explícitas. Essas regras são compostas por literais (i.e., palavras ou suas negações) e representam padrões lógicos de decisão. A principal contribuição do artigo é a demonstração de que cláusulas compostas majoritariamente por literais negados são menos sensíveis a correlações espúrias. Os autores introduzem o parâmetro de *especificidade* s, que controla a granularidade da cláusula: valores menores de s favorecem a inclusão de termos negados. Essa configuração permite que o modelo identifique padrões mais gerais e evite associações espúrias com palavras específicas. O modelo resultante é capaz de classificar corretamente dados modificados (contrafactuais), mantendo alta acurácia sem depender de dados adicionais. Os experimentos foram conduzidos com o conjunto de dados IMDB, contendo resenhas com versões manuais de sentido invertido. A TM, com $s = 2$, obteve até 73,56% de acurácia nos dados contrafactuais, superando modelos como *Naive Bayes*, SVM, Bi-LSTM e ELMo, e apresentando desempenho competitivo com o BERT, mesmo sem pré-treinamento. O artigo ora analisado e nossa Tese compartilham o foco na detecção de correlações espúrias em classificação de texto, mas divergem na abordagem. O artigo propõe a TM, que aprende regras lógicas com negações controladas por um parâmetro manual (s), mostrando que cláusulas negadas são mais robustas a dados contrafactuais. Já a Tese propõe um método automatizado, agnóstico ao modelo, que combina explicações via XAI com clusterização não supervisionada para detectar padrões espúrios, inclusive compostos, com base em métricas de frequência e impacto da perturbação.

Além dos trabalhos detalhados, outros estudos exploram contrafactuais para avaliar e mitigar correlações espúrias. Veitch et al. 2021 introduzem o conceito de invariância contrafactual como formalização da robustez frente a alterações irrelevantes nos dados de entrada. Por meio de uma análise causal estruturada, os autores mostram que a capacidade de um modelo de manter suas predições inalteradas sob intervenções contrafactuais depende do grafo causal subjacente, exigindo, por exemplo, diferentes regularizações conforme a direção causal (causal ou anticausal) entre atributos e rótulos. Em contraste com nossa proposta, que infere padrões espúrios a partir de instâncias reais e medidas observáveis de erro e frequência, o trabalho ora analisado pressupõe conhecimento da estrutura causal, uma limitação prática em tarefas de PLN aplicadas a dados reais com múltiplas fontes de viés.

Por sua vez, Liu et al. 2022 avaliam a robustez de modelos de geração de linguagem natural frente a dados contrafactuais construídos manualmente, revelando que modelos de ponta como GPT-2 e T5 exploram correlações espúrias entre cabeçalhos de tabelas e operadores lógicos. Como resposta, propõem um codificador estrutural de fórmulas lógicas e treinamento com dados contrafactuais gerados automaticamente. Embora eficazes, tais abordagens exigem pré-processamentos especializados e engenharia de atributos específicos da tarefa. Em contraste, nossa proposta aplica técnicas de explicabilidade e aprendizado não supervisionado de forma automatizada e agnóstica ao domínio, permitindo a detecção de padrões compostos espúrios sem necessidade de conhecimento causal prévio nem construção de contrafactuais.

3.3 Perturbação de Dados

Os trabalhos apresentados nesta seção aplicam perturbações locais nos dados de entrada — como inserção, substituição ou remoção de palavras — para estimar a influência de atributos nas decisões do modelo. Diferente da geração de contrafactuais, essas alterações não criam novas sentenças, focando no impacto direto das modificações sobre a predição. Com isso, é possível isolar a contribuição de cada termo e identificar aqueles que afetam o modelo.

Wang e Culotta 2020b tratam o problema de correlações espúrias como uma tarefa de classificação binária de palavras. O objetivo é distinguir entre correlações genuínas e espúrias. A abordagem parte do treinamento de um classificador inicial (regressão logística), utilizado para identificar as palavras mais associadas a cada classe com base nos coeficientes do modelo. Em seguida, para cada palavra, são construídos pares de sentenças semelhantes com e sem a palavra-alvo, utilizando medidas de similaridade baseadas em *embeddings* do BERT e similaridade de cosseno. Esse processo visa estimar o efeito de tratamento médio (*Average Treatment Effect* - ATE), uma métrica derivada da inferência causal que indica o quanto a presença da palavra altera a classificação em relação a sentenças comparáveis. A partir dessas comparações, são extraídas variáveis para caracterizar cada palavra, incluindo similaridade de contexto, variação nos *embeddings* e magnitude do ATE. Com um conjunto rotulado manualmente (cerca de 200 a 300 palavras classificadas como espúrias ou genuínas), é treinado um segundo classificador que aprende a identificar automaticamente a espuriedade de outras palavras. O classificador de palavras pode ser aplicado em diferentes domínios sem perda significativa de desempenho, o que demonstra a portabilidade do método. Por fim, os autores utilizam as probabilidades atribuídas pelo classificador de palavras para informar a seleção de atributos no classificador original, removendo palavras com alta chance de serem espúrias e avaliando o impacto dessa remoção em subconjuntos de dados mais sensíveis à espuriedade.

Embora apresente resultados promissores, o método exige intervenção manual para rotular palavras durante a fase de treinamento do classificador de palavras, além de operar sobre palavras individuais, o que limita a captura de padrões compostos. Em contraste, nossa proposta não depende de intervenção manual nem de rótulos manuais em nenhuma etapa. Além disso, enquanto o trabalho ora analisado realiza uma classificação binária dos padrões, nossa abordagem adota uma métrica contínua de espuriedade, o que permite uma avaliação gradual do potencial espúrio. Outra diferença relevante é que nossa proposta identifica padrões compostos, ampliando a capacidade de análise de dependências complexas entre *tokens*.

Wang et al. 2021 partem da constatação de que modelos modernos de PLN, especialmente redes neurais profundas, tendem a aprender padrões espúrios. Para lidar com isso, propõem um método para identificar esses padrões em larga escala e aplicar técnicas de mitigação. O método proposto é composto por três etapas principais. A primeira etapa extrai automaticamente *tokens* relevantes para a decisão do modelo, utilizando técnicas de interpretabilidade, como *attention scores* de modelos BERT e *Integrated Gradients*. O objetivo é identificar palavras que mais influenciam a predição, gerando um conjunto de *tokens* importantes com pesos atribuídos. Na segunda etapa, realiza-se uma análise de estabilidade entre domínios para distinguir *tokens* genuínos de espúrios. *Tokens* relevantes em múltiplos domínios são genuínos, enquanto *tokens* com alta importância em apenas um domínio são considerados espúrios. A última etapa envolve a substituição semântica dos *tokens* por sinônimos extraídos de bases como WordNet e DBpedia. Se a substituição afetar significativamente a predição, o *token* é classificado como espúrio devido ao impacto instável na decisão do modelo. O trabalho propõe três formas de mitigação: exclusão dos padrões espúrios no treinamento, remoção durante a inferência, ou combinação de ambas. Os experimentos mostram que essas estratégias aumentam a robustez e a equidade dos modelos, melhorando o desempenho em domínios fora da distribuição e tarefas sensíveis à justiça. Embora eficaz, a abordagem depende de intervenções humanas e pré-definições, como a necessidade de conjuntos de dados rotulados para análise entre domínios e a utilização de bases externas (WordNet, DBpedia) para substituição semântica. A validação dos padrões espúrios também requer anotadores humanos.

Em comparação com nossa proposta, o trabalho ora analisado foca em *tokens* individuais e valida a espuriedade com base na instabilidade à substituição semântica e divergência entre domínios e não define uma métrica de espuriedade do *token*, enquanto nossa abordagem considera padrões simples e compostos extraídos de sentenças erradas e utiliza métricas de impacto nas métricas de desempenho.

O artigo de Joshi, Pan e He 2022 propõe uma abordagem baseada em inferência causal para classificar diferentes tipos de correlações espúrias em tarefas de (PLN). Os autores distinguem entre padrões irrelevantes — cuja presença não altera a predição do modelo — e padrões necessários, que embora não sejam suficientes sozinhos, são essenciais para a decisão correta quando combinados ao contexto. Para isso, o trabalho introduz duas métricas: a probabilidade de necessidade (PN), que estima se a remoção de um termo muda a predição, e a probabilidade de suficiência (PS), que estima se sua adição é suficiente para provocar a predição esperada. Ambas são calculadas com base em perturbações simuladas por modelos de linguagem mascarados. A análise é aplicada a tarefas de inferência textual, como MNLI e HANS, avaliando como o modelo RoBERTa reage à presença de palavras como conectivos ou termos de negação. Os autores mostram que técnicas como balanceamento de dados ou remoção de atributos são eficazes para padrões irrelevantes, mas podem prejudicar o desempenho quando aplicadas a padrões de alta PN, que exigem composição contextual para gerar a predição correta.

Comparando com nossa proposta, ambas as abordagens compartilham o uso de perturbações nos dados para medir o impacto de padrões na decisão do modelo. No entanto, o método ora analisado exige intervenção manual em etapas-chave, como a definição dos termos a serem analisados e a interpretação da causalidade com base em conhecimento de domínio.

O artigo de Chang et al. 2022 propõe um método baseado em inferência causal para identificar correlações espúrias em classificadores de texto. O núcleo do método é o cálculo do Efeito de Tratamento Individual (*Individual Treatment Effect* - ITE), que estima o impacto causal de uma palavra sobre a predição de uma sentença. Três abordagens são utilizadas para estimar o ITE: emparelhamento de sentenças semelhantes, regressão logística e *perceptron* multicamada (MLP). A ideia central é avaliar como a remoção de uma palavra altera a previsão de sentimento de uma sentença. Se a remoção não afeta significativamente a predição, a palavra é considerada espúria. Para treinar o classificador de palavras espúrias, o artigo também considera diversas métricas adicionais baseadas em similaridade de contexto, influência média,

e diferença vetorial em *embeddings*. O método é aplicado nas bases IMDB e Kindle, e demonstra resultados superiores à remoção aleatória de palavras, especialmente no grupo majoritário, onde a correlação espúria é mais prevalente. Contudo, há queda de desempenho no grupo minoritário. A abordagem depende da definição e estimativa explícita de ITE para cada palavra, o que exige pré-configuração humana para seleção de pares de sentenças e definição de similaridade.

O método proposto por Serrano, Dodge e Smith 2023 investiga as correlações espúrias entre palavras e rótulos em tarefas de PLN. Para isso, os autores introduzem uma técnica estatística baseada em teste de perturbação que avalia se modelos treinados em dados enviesados apresentam maior acurácia em instâncias com rótulos "esperados" — ou seja, instâncias em que a presença de determinadas palavras coincide com a classe mais frequente — do que em instâncias com rótulos incomuns. Essa diferença é interpretada como sinal de viés aprendido. Após confirmar esse viés em modelos pré-treinados como RoBERTa-large e também em LSTMs treinados do zero, os autores aplicam uma técnica de reponderação dos dados. Essa técnica consiste em ajustar os pesos das instâncias de treino de forma a reduzir associações espúrias. Embora a reponderação reduza significativamente o viés nos dados, os modelos continuam a apresentar viés após o treinamento, sobretudo quando se considera padrões mais complexos (bigrams), que chegam a se tornar ainda mais enviesados. Os autores concluem que o viés lexical é persistente e que a mitigação puramente via dados é insuficiente. O método exige intervenção manual, pois depende da seleção *tokens* frequentes e da definição de um alvo de distribuição balanceada.

Comparado à nossa proposta, o trabalho ora analisado exige intervenção prévia no conjunto de dados, com foco em unigramas, e usa uma abordagem determinística e supervisionada para reponderar instâncias. Nossa proposta aplica (XAI) para identificar padrões importantes, realiza perturbação nos dados para medir impacto direto de cada padrão nas métricas do modelo sem intervenções.

Outras propostas utilizam perturbações de dados para investigar padrões espúrios. Plumb, Ribeiro e Talwalkar 2021 apresentam o método Identificação e reparo de padrões espúrios (*Spurious Pattern Identification and REpair* - SPIRE), voltado à identificação e mitigação de padrões espúrios em tarefas de classificação de imagens. A proposta parte da manipulação de imagens reais por meio da remoção de objetos anotados e observando a mudança nas predições do modelo. Se a remoção de um objeto altera significativamente a predição, o padrão é considerado espúrio. O método requer anotações pixel a pixel, intervenção manual para validar os padrões identificados, e uma infraestrutura de manipulação de imagens. Como forma de mitigação, aplica estratégias de aumento de dados para balancear a presença de padrões espúrios nos dados de treino.

O trabalho ora analisado compartilha com nossa proposta o objetivo central de identificar padrões espúrios por meio de intervenções sistemáticas nas entradas e avaliar seu impacto nas decisões do modelo. Ambos os métodos operam com a hipótese de que alterações em partes específicas da entrada que causam mudanças na predição revelam dependências espúrias. Enquanto SPIRE remove objetos anotados em imagens, nossa abordagem remove padrões textuais identificados como importantes e observa variações nas métricas de desempenho. Em comum, as duas técnicas combinam explicabilidade e modificação do dado para quantificar espuriedade, avaliam o efeito da remoção de padrões em métricas de desempenho específicas e utilizam esses efeitos para orientar a análise. A principal diferença está no domínio (visão computacional versus PLN) e na necessidade de intervenção manual, ausente na nossa abordagem.

Já, Ming, Yin e Li 2022 propõem uma reformulação do problema de detecção fora da distribuição (*Out-of-Distribution* – OOD), baseada na separação entre características relevantes ao rótulo. O estudo investiga dois tipos de exemplos OOD — espúrios e não espúrios — e avalia o impacto da correlação espúria nos dados de treinamento sobre métodos clássicos de detecção OOD. Os autores demonstram que modelos treinados com risco empírico minimizado podem aprender dependências espúrias mesmo quando há invariância entre

domínios. A abordagem, contudo, depende de conhecimento prévio sobre as características ambientais e é limitada a domínios visuais. De forma híbrida, Shirnin et al. 2024 exploram a robustez de modelos multimodais de Visão e Linguagem (V&L) na tarefa de respostas visuais a perguntas (*Visual Question Answering* - VQA), por meio da aplicação sistemática de perturbações em textos e imagens. Os autores avaliam modelos como LXMERT, VisualBERT e OFA, aplicando alterações como erros ortográficos e distorções visuais, além de analisar a sensibilidade dos modelos à remoção de palavras e ao desalinhamento entre modalidades, utilizando mapas de atenção. A proposta é diagnóstica e baseada em abordagem *black-box*, com múltiplas intervenções definidas manualmente. Embora revele padrões espúrios, a análise é restrita a tarefas multimodais e requer interpretação qualitativa dos resultados.

No domínio textual, Zhang et al. 2021 introduzem o conceito de *learnability* para medir a sensibilidade de modelos de PLN a diferentes perturbações artificiais. O método estima o quão facilmente o modelo distingue exemplos perturbados dos originais, sendo aplicada em cenários controlados com perturbações como erros de digitação, duplicações de pontuação e substituições visuais. Os experimentos envolvem modelos como BERT, RoBERTa e XLNet. A proposta evidencia que perturbações mais aprendíveis estão associadas a maior impacto negativo sobre a robustez, mas exige pré-definição das manipulações e não contempla padrões compostos. Já Wu e Hooi 2022 analisam correlações espúrias em *benchmarks* populares de detecção de boatos. Os autores identificam fontes de viés ligadas à coleta por evento e ao estilo de escrita de publicadores, propondo o método Agregação de estilo do editor (*Publisher Style Aggregation)* - PSA), que agrupa múltiplas postagens de um mesmo autor para construir representações de estilo textual. Essa representação é combinada com a do *microblog* individual para prever a veracidade. A proposta melhora a generalização entre eventos, mas requer múltiplas postagens por autor e associação explícita entre conteúdo e fonte, restringindo seu uso a mídias sociais.

3.4 Inteligência Artificial Explicável (XAI)

Esta seção reúne trabalhos que exploram, direta ou indiretamente, técnicas XAI para investigar a presença e o impacto de correlações espúrias. Em geral, esses métodos utilizam explicadores para identificar os atributos mais relevantes para as decisões do modelo, partindo do pressuposto de que padrões recorrentes entre os atributos explicativos podem indicar dependências espúrias. O uso de XAI é particularmente relevante no contexto desta Tese, pois assumimos que, como (Wang e Culotta 2020a, Wang e Culotta 2020b, Wang et al. 2021), atributos espúrios são frequentemente apontados como importantes pelos modelos, podendo ser revelados por meio de técnicas XAI.

O artigo de Jakobsen, Barrett e Søgaard 2021 propõe uma análise sobre correlações espúrias no contexto de mineração de argumentos entre tópicos distintos. O estudo avalia modelos de classificação de argumentos que, mesmo sendo treinados com protocolos *cross-topic* (avaliando em tópicos não vistos), continuam a depender de palavras específicas dos tópicos — consideradas correlações espúrias. O método envolve o uso de LIME para examinar os termos mais influentes nas decisões dos modelos. Além disso, realiza experimentos com ablações de vocabulário, categorização manual de palavras e modelos restritos a palavras de classes fechadas (como preposições e conjunções). Uma descoberta central é que modelos com acesso apenas a palavras de classe fechada e algumas poucas palavras de classe aberta compartilhadas entre tópicos têm desempenho superior na transferência entre tópicos distantes, indicando que a maioria dos modelos estuda padrões espúrios mesmo quando avaliados nesses cenários.

Em comparação com nossa proposta, há similaridades importantes: ambos os trabalhos investigam correlações espúrias com apoio de técnicas de explicabilidade e consideram que essas correlações afetam a generalização dos modelos. Ambos também realizam análises sobre as palavras mais relevantes para os modelos, tratando essas palavras como potenciais padrões espúrios. Contudo, há diferenças metodológicas. O trabalho ora analisado depende da categorização manual das palavras, consideradas importantes pelo LIME, como

argumentativas ou espúrias, exigindo intervenção manual tanto na classificação dos padrões quanto na análise qualitativa, limita-se a palavras individuais e necessitam de conhecimento prévio para definir o que é espúrio, nossa abordagem não requer predefinições humanas, é agnóstica ao modelo, e identifica padrões espúrios de forma automática com base em múltiplas métricas quantitativas.

Lampridis et al. 2023 apresenta o método XSPELLS (eXplaining Sentiment Prediction gEnerating exempLars in the Latent Space). A explicação consiste em sentenças sintéticas com rótulo igual (exemplares) ou diferente (contra-exemplares) ao do texto analisado, geradas a partir do *espaço latente*, uma representação vetorial compacta que preserva informações semânticas do texto. As sentenças são criadas por autocodificadores variacionais (*Variational Autoencoders - VAEs)*, que codificam textos em vetores e permitem gerar novas amostras semanticamente próximas. A partir desses vetores, uma árvore de decisão é treinada para aproximar localmente o comportamento do modelo original, e suas ramificações são usadas para selecionar exemplares e contra-exemplares mais relevantes. Palavras frequentes nessas sentenças também são destacadas para apoiar a interpretação.

Embora o foco do trabalho não seja a detecção de correlações espúrias, a ideia de contraste entre vizinhos com rótulos distintos inspirou a estratégia adotada nesta tese. Os autores testaram o *xspells* em cinco conjuntos de dados de classificação binária de textos curtos, incluindo sentimentos, discurso de ódio e detecção de spam. Os resultados mostraram que o método superou o LIME em fidelidade, utilidade, diversidade e compreensibilidade das explicações.

O artigo de Chew et al. 2024 propõe o método *doN't Forget your Language* (NFL), voltado à mitigação de correlações espúrias em tarefas de classificação de texto. O método se baseia em uma análise de vizinhança (*neighborhood analysis*), que verifica como o espaço de *embeddings* muda durante o *fine-tuning* de modelos baseados em *tranformers*. A ideia central é que *tokens* espúrios se aproximam indevidamente de *tokens* genuínos nesse espaço vetorial, adquirindo importância exagerada nas predições. Para evitar isso, os autores introduzem uma família de regularizações: NFL-F (congela o modelo),

NFL-CO (constrange as saídas do modelo), NFL-CP (penaliza alterações nos parâmetros) e NFL-PT (usa *prompt-tuning*). As regularizações restringem a deformação dos *embeddings* originais durante o treino. Os experimentos mostram que essas variantes do NFL mantêm baixa pontuação de espuriedade para *tokens* espúrios e melhoram a robustez dos modelos em bases como *Amazon Binary* e *Jigsaw*, com precisão robusta próxima ao modelo ideal treinado em dados livres de viés. Comparado à nossa proposta, ambos os trabalhos usam variações de predição e medidas internas para inferir espuriedade. No entanto, há diferenças relevantes: enquanto nossa proposta atua no nível de padrões compostos extraídos de erros do modelo, combinando explicações locais (LIME) com clusterização de métricas de perturbação, o método NFL atua no nível dos *tokens* individuais, focando na regularização do modelo e não na análise direta de erros. Além disso, o NFL requer intervenção no processo de treinamento via regularização, enquanto nossa abordagem analisa modelos já treinados, com foco na explicação e perturbação de padrões para detectar espuriedade. Por fim, o método ora analisado exige escolha prévia da técnica de regularização (NFL-F, NFL-CO etc.), configurando uma forma indireta de predefinição humana, o que não ocorre em nossa abordagem totalmente automática e agnóstica.

3.5 Outras Técnicas

Esta seção aborda propostas que adotam estratégias complementares para identificar e mitigar correlações espúrias. Essas abordagens introduzem arquiteturas específicas ou mecanismos de regularização que buscam melhorar a robustez dos modelos a padrões não causais. Em geral, combinam técnicas de aprendizado supervisionado com hipóteses estruturais ou estatísticas sobre o comportamento desejado dos modelos frente a dados espúrios.

O artigo de Yang et al. 2023 propõe o método representação causal para aprendizagem de poucas tentativas (*Causal Representation for Few-Shot Learning* - CRFL), aplicado à classificação de texto com poucos exemplos (few-shot text classification). O objetivo é extrair representações causais a partir do texto, separando fatores causais (S) dos não causais (U), para melhorar a generalização e evitar correlações espúrias. O método utiliza três módulos:

(i) intervenção causal, que gera versões sinônimas e antônimas dos textos para identificar representações invariantes (S); (ii) fatoração causal, que força a independência entre as dimensões das representações aprendidas; e (iii) classificação, que usa as representações para prever rótulos. Os experimentos foram realizados em seis base de dados, incluindo SST-2, MR e AGNews. O CRFL superou métodos anteriores, mostrando ganho médio de até 1,6% em acurácia, mesmo com menos exemplos de treino. Em relação à nossa proposta, a principal semelhança está no uso de intervenções nos dados para analisar a relação entre padrões e classes. Contudo, o CRFL requer pré-definições humanas, como o uso explícito de sinônimos e antônimos para perturbar os dados, além de uma arquitetura de rede neural específica com treinamento supervisionado.

O artigo de Ghosal e Li 2023 propõe o método otimização distribucional robusta de grupo probabilístico (*Probabilistic Group Distributionally Robust Optimization* - PG-DRO), um método de otimização robusta à distribuição que utiliza grupos probabilísticos em vez de rótulos fixos. Ao contrário de abordagens anteriores, que assumem que cada exemplo pertence a um único grupo, o PG-DRO admite incerteza na atribuição de grupos, representando cada exemplo como pertencente a múltiplos grupos com probabilidades associadas. O método é composto por duas etapas principais: (i) geração de rótulos probabilísticos por meio de pseudo-rotulagem, com uso de dados anotados; e (ii) otimização robusta baseada na minimização do risco do pior grupo, ponderado pelas probabilidades de grupo de cada exemplo. A abordagem é avaliada em *benchmarks* de visão computacional e PLN, e demonstrou desempenho superior em acurácia no pior grupo, mesmo usando apenas 5% dos dados com anotação de grupo.

Em comparação com nossa proposta, ambos os métodos são agnósticos ao modelo de classificação. No entanto, o PG-DRO requer uma etapa de pseudo-rotulagem para estimar as probabilidades de grupo, o que introduz dependência de heurísticas ou de um subconjunto anotado, configurando uma intervenção manual. Em contraste, nossa proposta não exige nenhuma informação sobre grupos e identifica padrões espúrios a partir de análise de erros, explicabilidade e clusterização, sem suposições sobre a estrutura dos dados. Além disso,

enquanto o PG-DRO atua na função de perda global com base em rótulos de grupo, nossa abordagem atua diretamente sobre os padrões linguísticos específicos responsáveis pelos erros, quantificando seu impacto por meio de perturbação no conjunto de teste.

Esses trabalhos ilustram diferentes abordagens para detecção de correlações espúrias. Observa-se que métodos baseados em geração de contrafactuais e perturbações de dados predominam na literatura, frequentemente combinados inferência causal e regularização. No entanto, a maioria dessas propostas apresenta limitações quanto à intervenção manual, exigência de pré-definições, detectam atributos simples (unigramas) e adotam métricas binárias (é/não é) para identificação de espuriedade.

A proposta desta Tese diferencia-se por combinar XAI com aprendizado não supervisionado em um método automático, capaz de detectar padrões espúrios compostos com base em evidências observáveis de erro, sem necessidade de contrafactuais ou estruturas causais explícitas. O uso da distância aos centroides como métrica contínua de espuriedade proporciona uma avaliação mais graduada e interpretável, aplicável a múltiplos modelos e domínios. Além disso, a extração dos padrões diretamente a partir de erros reais dos modelos permite uma análise mais alinhada com comportamentos efetivamente observados. A Tabela 1 apresenta uma síntese comparativa entre os trabalhos relacionados e a proposta desta Tese. A tabela explicita quatro dimensões centrais de comparação: se o método é agnóstico ao modelo, se identifica padrões compostos, se a detecção é totalmente automatizada e qual tipo de métrica é utilizada para quantificar a espuriedade.

Em síntese, a análise da literatura evidencia avanços importantes, mas também limitações recorrentes, como a dependência de conhecimento prévio, o foco em atributos simples e a ausência de métricas graduais de espuriedade. Essa lacuna reforça a necessidade de um método alternativo, capaz de operar de forma automática, agnóstica e orientada por métricas estatísticas e de impacto sobre o modelo. O próximo capítulo apresenta a proposta metodológica desenvolvida nesta Tese, detalhando suas etapas, bases de dados e estratégias de integração entre explicabilidade e aprendizado não supervisionado.

Tabela 1 – Comparação entre trabalhos relacionados e a proposta da Tese, com base nos critérios: agnosticismo ao modelo, reconhecimento de padrões compostos (PC), automação completa do processo e métrica adotada para quantificar espuriedade.

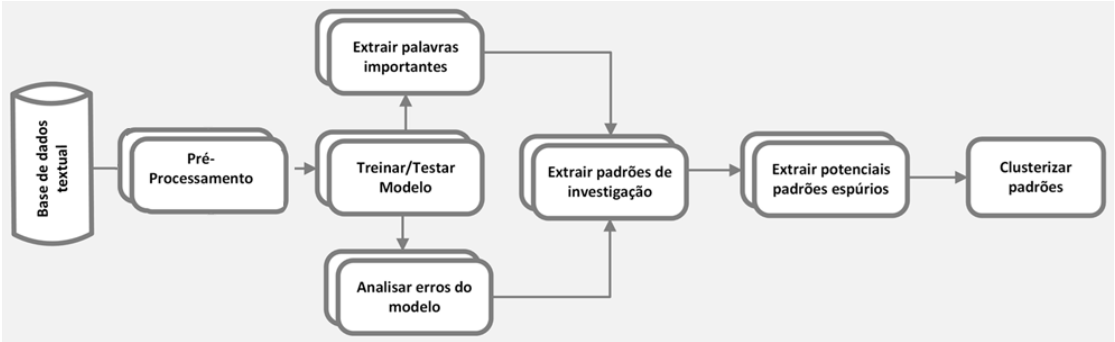
Trabalho	Técnica	Agnóstico	PC	Automático	Métrica
Contrafactuais					
Wang e Culotta 2020a	Substituição por antônimos com *matching* oposto.	-	-	-	-
Veitch et al. 2021	Regularização baseada em invariância contrafactual.	-	-	-	-
Wu et al. 2022	Geração com GPT-2 e filtragem por z-score.	Parcial	-	-	*z-score*
Yadav et al. 2022	Regras lógicas com negações via *Tsetlin Machine*.	Parcial	Sim	-	-
Liu et al. 2022	Contrafactuais estruturais em geração de linguagem.	-	-	-	-
Perturbação de dados					
Wang e Culotta 2020b	Pareamento com ATE para classificar palavras.	-	-	-	ATE
Wang et al. 2021	Substituição semântica e atenção para *tokens* espúrios.	-	-	-	-
Plumb, Ribeiro e Talwalkar 2021	SPIRE: remoção de objetos em imagem.	-	-	-	-
Zhang et al. 2021	Aprendibilidade por perturbações artificiais.	Parcial	-	-	*learnability*
Joshi, Pan e He 2022	PN/PS com máscaras para suficiência e necessidade.	-	-	-	PN/PS
Chang et al. 2022	Cálculo de ITE por MLP e pareamento.	-	-	-	ITE
Ming, Yin e Li 2022	Separação de atributos em OOD.	-	-	-	-
Serrano, Dodge e Smith 2023	Reponderação estatística de instâncias enviesadas.	-	-	-	-
Wu e Hooi 2022	Agregação de estilo de autor (PSA).	-	-	-	-
Shirnin et al. 2024	Perturbações em VQA multimodal.	-	-	-	-
Explicabilidade (XAI)					
Jakobsen, Barrett e Søgaard 2021	LIME com análise manual de importância.	Parcial	-	-	-
Lampridis et al. 2023	XSPELLS: exemplares e contra-exemplares sintéticos.	Sim	-	Sim	-
Chew et al. 2024	Regularizações NFL com análise de *embeddings*.	-	-	-	-
Outros					
Yang et al. 2023	Representação causal com sinônimos e antônimos.	-	-	-	-
Ghosal e Li 2023	PG-DRO: otimização robusta com grupos probabilísticos.	Sim	-	-	-
Nossa proposta	Clusterização de padrões com XAI e perturbação.	Sim	Sim	Sim	Distância ao centroide

4 O método proposto

Este capítulo apresenta o método proposto para detectar padrões espúrios em modelos de PLN. Inicialmente, serão descritas as bases de dados utilizadas. Em seguida, será detalhado cada etapa do processo e suas respectivas contribuições.

O método proposto organiza-se em sete etapas: Pré-processamento, Treinar/Testar modelo, Extrair palavras importantes, Analisar erros, Extrair de padrões de investigação, Extrair potenciais padrões espúrios e Clusterizar padrões, conforme ilustrado na Figura 1. As etapas representadas com linhas duplicadas indicam que foram executadas sobre uma configuração da base de dados, obtida pela materialização da validação cruzada (*k-fold cross validation*, comumente utilizada em PLN (Kohavi et al. 1995). Neste estudo adotamos $k = 5$, mas o número de partições pode ser ajustado conforme a necessidade experimental. As técnicas que podem ser aplicadas a cada etapa variam conforme o modelo, o que torna nossa proposta agnóstica.

Figura 1 – O Método: Pré-processamento; Treinar/Testar Modelo; Extrair palavras importantes; Analisar Erros; Extrair padrões de investigação; Extrair padrões espúrios; e Clusterizar padrões.



Cada etapa do método será explicada nas seções a seguir, as definições e notações adotadas seguem as convenções estabelecidas nos trabalhos de Wang e Culotta 2020b e Pezeshkpour et al. 2021.

4.1 Bases de Dados

Neste trabalho, utilizamos bases de dados do contexto do Tribunal de Contas do Estado do Piauí (TCE-PI), denominados (i) Base de dados de Contratos (C) e (ii) base de dados de Licitações (L), que foram desenvolvidas a partir de pesquisas produzidas no Programa de Pós-Graduação em Ciência da Computação na Universidade Federal do Piauí (PPGCC/UFPI) e do curso de Doutorado em Ciência da Computação Associação UFMA - UFPI (DCCMAPI), respectivamente.

A base de dados de Contratos foi elaborada para desenvolver um modelo de classificação automática de contratos da Administração Pública, com foco em gastos relacionados à COVID-19 (Vale et al. 2023). A rotulagem foi realizada por 12 auditores especialistas do TCE-PI, que analisaram descrições de objetos de contratos, extraídas dos Diários Oficiais dos Municípios e do Estado do Piauí referentes ao período de março a setembro de 2020. Cada auditor avaliou um conjunto de descrições, com revisões por amostragem realizadas por outro auditor e, em caso de divergência, a decisão final ficou a cargo do auditor-chefe. As descrições foram categorizadas em duas classes: aquisições específicas da saúde (Classe 1) e demais aquisições (Classe 0).

Para este estudo, a equipe expandiu o conteúdo com objetos de contratos extraídos do Sistema Contratos-Web[1], uma plataforma do TCE-PI onde entidades jurisdicionadas registram contratos conforme exigido por lei. Os especialistas aplicaram o mesmo processo de rotulagem a essas novas entradas, adicionando 1.727 novos objetos de contrato. A tabela 2 resume as quantidades da base de dados antes e depois da expansão.

Tabela 2 – Resumo da base de dados de contratos original e expandida.

Classe	Rótulo	Quantidade Inicial	Quantidade Após Expansão
Aquisições específicas da saúde	1	2.067	2.918
Demais aquisições	0	2.298	3.174
Total		**4.365**	**6.092**

[1] <https://sistemas.tce.pi.gov.br/contratosweb/>, acessado em 19 de março de 2025.

O conjuntos de dados de Licitações foi inicialmente desenvolvido para avaliar uma abordagem arquitetônica que identifique indícios de fraude em licitações públicas, conforme descrito em Lima et al. 2025. Consiste em editais de licitações publicados entre 2012 e 2023, registrados no sistema Licitações-Web[2] do TCE-PI. O sistema permite que as entidades jurisdicionais registrem seus contratos conforme exigido pela legislação vigente. O processo de extração de dados utilizou expressões regulares para isolar a seção que descreve o objeto licitado. Para verificar a precisão da extração, treinamos um modelo BERT especificamente para essa tarefa. Uma comissão de três especialistas em licitações públicas rotulou manualmente os dados em quatro categorias principais. A tabela 3 mostra as quantidades de dados rotulados para cada classe:

Tabela 3 – Quantidade de dados rotulados por classe e seus respectivos rótulos.

Classe	**Rótulo**	**Quantidade**
Contratação de serviços de obras de engenharia	0	650
Aquisição de bens permanentes	1	370
Aquisição de bens de consumo	2	623
Contratação de serviços em geral	3	494
Total		**2.137**

Para este estudo, expandimos a base de dados original usando o *Active Learning* (Cohn, Atlas e Ladner 1994) com adaptações específicas para classificação binária. Inicialmente, isolamos apenas as sentenças que descrevem o objeto licitado conforme as instruções a seguir dos especialistas. Sentenças que apresentam padrões sintáticos como “verbo + preposição” ou “substantivo (derivado de verbo) + preposição”. Para determinar a classe gramatical das palavras, utilizamos *Part-of-Speech Tagging* (POS Tagging) com um modelo baseado em (Sousa e Lopes 2019), que emprega uma arquitetura BLSTM e a última versão do corpus Mac-Morpho (Fonseca e Rosa 2013). Agrupamos as classes “Contratação de serviços de obras de engenharia” e “Contratação de serviços em geral” na nova classe 0; e as classes “Aquisição de Bens Permanentes” e “Aquisição de Bens de Consumo” na nova classe 1.

[2] <https://sistemas.tce.pi.gov.br/licitacoesweb/>, acessado em 19 de março de 2025

Para rotular a as sentenças, utilizamos a base de dados já rotulada e adaptada para treinar um modelo BERT. Então, selecionamos, de forma aleatória, 200 sentenças ainda não classificadas e aplicamos o modelo BERT para classificá-las. Os especialistas revisavam as classificações, corrigindo rótulos incorretos ou descartando sentenças que não correspondessem ao objeto licitado. As sentenças corrigidas eram então adicionadas à base de treinamento e iniciada mais uma rodada de classificação/revisão. O processo foi repetido até que a quantidade de sentenças em cada nova classe se aproximou daquelas disponíveis na base de Contratos. A tabela 4 mostra a distribuição final das sentenças após a aplicação do método de expansão da da base de dados:

Tabela 4 – Resumo da base de dados de licitações original e expandida.

Classe	**Rótulo**	**Quantidade Inicial**	**Quantidade Após Expansão**
Aquisição de bens	1	993	3.040
Contratação de serviços	0	1.144	3.204
Total		**2.137**	**6.244**

Definimos a base de dados da seguinte forma. Seja $W = \{w_1, w_2, ..., w_n\}$ o conjunto de palavras e símbolos em uma língua, e seja $S \subseteq W^+$ o conjunto de sentenças válidas, onde W^+ representa o conjunto de todas as sequências finitas não vazias de elementos de W. A base de dados é particionada em conjuntos de treinamento e teste disjuntos, $S_{train} \subset S$ e $S_{test} \subset S$, garantindo $S_{train} \cap S_{test} = \emptyset$. A tarefa de classificação é definida sobre o conjunto de rótulos $Y = \{0, 1\}$, onde cada sentença recebe um rótulo exclusivo. Os conjuntos de dados de treinamento e teste são então fornecidos por $D_{train} = \{(s_r, y) \mid s_r \in S_{train}, y \in Y\}$ e $D_{test} = \{(s_t, y) \mid s_t \in S_{test}, y \in Y\}$, onde cada sentença de treino s_r, ou de teste s_t, é associada exclusivamente a um único rótulo y.

Como já explicado anteriormente, materializamos um *cross-validation* cinco *folds* com D_{train} e D_{test} independentes, essa abordagem reduz o viés da estimativa e a variância dos resultados, conforme discutido em (Kohavi et al. 1995, Hastie et al. 2009). Esses *folds* funcionam como partições fixas da base de dados, garantindo que cada instância participasse do treinamento e do teste em diferentes contextos.

4.2 Pré-processamento

O pré-processamento tem como objetivo transformar as sentenças textuais em representações estruturadas adequadas à etapa de modelagem. De forma geral, esse processo compreende a normalização dos textos, a identificação das unidades linguísticas mais relevantes e a conversão dessas unidades em representações numéricas compatíveis com os modelos utilizados. A forma como essa transformação é realizada pode variar conforme as características do modelo, mas segue o princípio comum de preparar os dados para maximizar a eficiência do aprendizado. Além disso, são adotadas estratégias para garantir que as informações contextuais e semânticas sejam preservadas ou realçadas, quando possível, a fim de melhorar a capacidade dos modelos de capturar relações significativas nos dados.

4.3 Treinar/Testar o modelo

Esta etapa define o processo de treinamento e teste dos modelos. Para cada partição da base de dados, aplicamos validação cruzada *k-fold*, com $k = 5$. Definimos o modelo treinado nos dados de treinamento D_{train}, denotado por:

$$\mathsf{F}_{(g)} = f(g(D_{\text{train}})) \tag{4.1}$$

onde $g : s \mapsto x$ converte uma sentença s em uma representação vetorial x. A Equação 4.1 define o modelo de classificação como a composição de duas funções: $g : s \mapsto x$, que transforma a sentença s em uma representação vetorial x, e $f : x \mapsto y$, que realiza a predição da classe y. Assim, o modelo $\mathsf{F}_{(g)}$ corresponde à aplicação de f sobre as representações geradas por g a partir dos dados de treino D_{train}. A função g pode representar diferentes estratégias de codificação textual (e.g., TF-IDF ou *embeddings*), enquanto f pode ser um classificador linear ou uma rede neural. Em modelos lineares, apenas f é ajustado; em arquiteturas como BERT, f e g são otimizados conjuntamente.

Outra tarefa nesta etapa cataloga os erros do modelo definindo D_{erro} como o conjunto de tuplas de classe de sentenças do conjunto de teste onde o modelo fez previsões incorretas, conforme mostrado na seguinte notação:

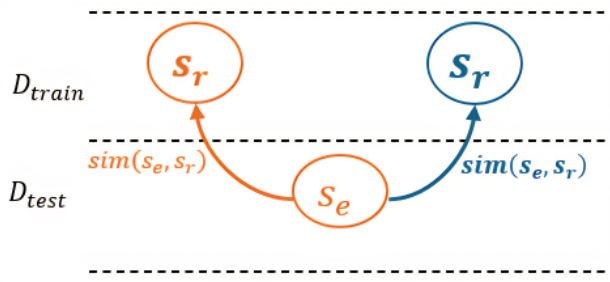
$$D_{erro} = \{(s_e, y_i) \mid (s_e, y_i) \in D_{test},\ y_i \neq \mathsf{F}_{(g)}[s_e]\}. \tag{4.2}$$

onde, s_e representa a sentença com erro, y_i é sua classe real e $\mathsf{F}_{(g)}[s_e]$ é a previsão feita pelo modelo treinado com D_{train} e aplicado à sentença s_e. Estas definições são essenciais para a próxima etapa do método, pois permite uma análise detalhada dos erros do modelo. Além disso, D_{erro} serve como base para a identificação de candidatos a espúrios, contendo padrões comuns nas sentenças erroneamente classificadas durante e teste.

4.4 Analisar erros do modelo

Sentenças semelhantes, em D_{train}, com rótulos diferentes, induzem o aprendizado de associações incorretas, elevando a taxa de erro (Zhou et al. 2023, Wu et al. 2023). Essa etapa analisa as sentenças classificadas incorretamente pelo modelo — chamadas de sentenças de erro (s_e) — e suas sentenças semelhantes, definidas como sentenças do conjunto de treinamento (s_r) semanticamente próximas a s_e. Chamamos de semelhantes da mesma classe aquelas com o mesmo rótulo de s_e, e de semelhantes da classe oposta aquelas com rótulo contrário. A figura 2 ilustra a relação, o objetivo é identificar padrões comuns entre as sentenças, buscando elementos que possam estar associados a correlações espúrias aprendidas pelo modelo.

Figura 2 – Relação entre sentenças de erro s_e e sentenças semelhantes s_r; sim(s_e, s_r) indica o grau de similaridade entre elas.



O Algoritmo 1 apresenta os passos necessários para identificar sentenças semelhantes às sentenças de erro. O algoritmo recebe como entrada o conjunto de treinamento D_{train} e o conjunto de classes $Y = \{0, 1\}$. Inicialmente, a estrutura de saída R é criada como uma lista com dois conjuntos vazios: R_0 e R_1.

Algorithm 1 Selecionar sentenças semelhantes.

Require: D_{train}: Conjunto de treinamento, Y: Conjunto de rótulos das classes
Ensure: R, explicado a seguir.

$R \leftarrow [\{\}, \{\}]$ // Um conjunto para cada classe
for $y \in Y$ **do**
  $D_{\text{erro}} \leftarrow \{(s_e, y_i) \mid (s_e, y_i) \in D_{test}, y_i \neq \mathsf{F}_{(g)}[s_e]\}$
  for $s_e \in D_{\text{erro}}$ **do**
    $\mathcal{M}_{(se)} \leftarrow \{\}$
    $\mathcal{O}_{(se)} \leftarrow \{\}$
    $sentencas_{\text{semelhantes}} \leftarrow \{(s_r, sim(s_e, s_r)) \mid s_r \in D_{\text{train}}\}$ // Calcula a similaridade entre s_e e todas as sentenças de D_{train}.
    while $\mathcal{M}_{(se)} = \{\} \wedge \lambda_m > 0$ **do**
      for $(sentenca, similaridade) \in sentencas_{\text{semelhantes}}$ **do**
        if $similaridade \geq \lambda_m \wedge \text{classe}(sentenca) = y$ **then**
          $\mathcal{M}_{(se)} \leftarrow \mathcal{M}_{(se)} \cup \{(sentenca, similaridade)\}$
        end if
      end for
      $\lambda_m \leftarrow \lambda_m - \delta_m$
    end while
    while $\mathcal{O}_{(se)} = \{\} \wedge \lambda_o > 0$ **do**
      for $(sentenca, similaridade) \in sentencas_{\text{semelhantes}}$ **do**
        if $similaridade \geq \lambda_o \wedge \text{classe}(sentenca) \neq y$ **then**
          $\mathcal{O}_{(se)} \leftarrow \mathcal{O}_{(se)} \cup \{(sentenca, similaridade)\}$
        end if
      end for
      $\lambda_o \leftarrow \lambda_o - \delta_o$
    end while
    $R[y] \leftarrow R[y] \cup \{(s_e, (\mathcal{M}_{(se)}, \mathcal{O}_{(se)}))\}$
  end for
end for
return R

Para cada classe y, o algoritmo identifica as sentenças de erro s_e , em seguida, para cada s_e, são criados dois conjuntos vazios: $\mathcal{M}_{(s_e)}$, destinado às

sentenças semelhantes de mesma classe, e $\mathcal{O}_{(s_e)}$, destinado às sentenças semelhantes da classe oposta. A variável `sentencas_semelhantes` é então construída como uma lista contendo pares $(s_r, sim(s_e, s_r))$, onde s_r é uma sentença de D_{train} e $sim(s_e, s_r)$ representa uma função que calcula o grau de similaridade entre s_e e s_r.

Para encontrar as sentenças semelhantes da mesma classe, o algoritmo percorre os pares em `sentencas_semelhantes` verificando se a similaridade é maior ou igual a um limiar λ_m e se a classe de s_r é igual a y. Se sim, o par (s_r, sim) é adicionado a $\mathcal{M}_{(s_e)}$. Esse processo se repete enquanto $\mathcal{M}_{(s_e)}$ estiver vazio e $\lambda_m > 0$, sendo que o limiar é reduzido progressivamente por um valor δ_m a cada iteração. O mesmo procedimento é aplicado para encontrar sentenças da classe oposta, utilizando o limiar λ_o e o decremento δ_o, preenchendo assim o conjunto $\mathcal{O}_{(s_e)}$. Por fim, a sentença de erro s_e, junto com seus conjuntos $\mathcal{M}_{(s_e)}$ e $\mathcal{O}_{(s_e)}$, é adicionada a R_y. O resultado final é representado por uma estrutura $R = [R_0, R_1]$, onde cada elemento R_y, $y \in Y = \{0, 1\}$, contém tuplas com uma sentença de erro s_e e dois conjuntos: um de sentenças semelhantes da mesma classe $\mathcal{M}(s_e)$ e outro da classe oposta $\mathcal{O}(s_e)$:

$$R_y = \{(s_e, (\mathcal{M}(s_e), \mathcal{O}(s_e))) \mid s_e \in D_{\text{erro}}\} \tag{4.3}$$

Cada conjunto associado a s_e é definido por:

$$\mathcal{M}(s_e) = \{(s_r, sim(s_e, s_r)) \mid (s_r, y_r) \in D_{\text{train}},\ y_r = y,\ sim(s_e, s_r) \geq \lambda_m\} \tag{4.4}$$

$$\mathcal{O}(s_e) = \{(s_r, sim(s_e, s_r)) \mid (s_r, y_r) \in D_{\text{train}},\ y_r \neq y,\ sim(s_e, s_r) \geq \lambda_o\} \tag{4.5}$$

onde:

- $sim(s_e, s_r)$: Corresponde à métrica de similaridade entre s_e e s_r.
- λ_m, λ_o: Limiares mínimos de similaridade para seleção de sentenças da mesma classe e da classe oposta, respectivamente.

Essa estrutura facilita a análise dos padrões que influenciam os erros do modelo. No contexto deste trabalho, essa seleção é útil para: (*i*) entender como o modelo agrupa sentenças semanticamente próximas; (*ii*) detectar padrões ambíguos entre classes que podem estar associados a erros de classificação.

4.5 Extrair palavras importantes

O objetivo desta etapa é identificar as palavras consideradas importantes pelos modelos. Neste trabalho, aplicamos técnicas de explicabilidade para extrair os atributos mais relevantes de cada classe e estabelecemos os requisitos necessários para os explicadores, como mostrado a seguir:

$$E_{xp}(\mathsf{F}_{(g)}, s) \mapsto E_{L(n)} = \{e_l\}, s \in S, e_l = (w_l, p_l) \in W \times \mathbb{R} \tag{4.6}$$

onde, E_{xp} representa um explicador que, dado um modelo $\mathsf{F}_{(g)}$ e uma sentença s, gera conjuntos de pares ordenados $E_{L(n)}$. Cada par ordenado e_l é composto por um *token* w_l e um peso p_l (um número real). O pesop_l reflete a importância de w_l para a classe n, permitindo compreender como o modelo $\mathsf{F}_{(g)}$ chegou à sua previsão ao identificar os *tokens* mais relevantes. Para explicadores intrínsecos, E_{xp} representa o mecanismo interno de explicação, como por exemplo em modelos baseados em Regressão Logística, o cálculo de coeficientes (pesos) para cada variável independente, que indicam a força e a direção da associação entre a característica e a resposta (Ali et al. 2023). Este estudo explora a classificação binária, portanto, n pode assumir apenas dois valores: 0 ou 1. Assim, para uma sentença s, são gerados dois conjuntos diferentes: $E_{L(0)}$ contendo os pares ordenados de w_i e p_l para a classe 0, e $E_{L(1)}$ contendo os pares ordenados de w_i e p_l para a classe 1, na sentença s. Essa abordagem permite uma análise detalhada e interpretável das características relevantes para cada classe. Essa abordagem é semelhante à explicação local. Para alcançar o objetivo desta etapa, é necessário calcular a importância global de cada *token*. Portanto, definimos a explicação global como apresentado na Equação 4.7:

$$E_{G(n)} = \{e_g\}, e_g = (w_g, p_g) \in W \times \mathbb{R} \tag{4.7}$$

onde, p_g representa o peso global do *token* w_g para a classe n. Aqui, também serão gerados dois conjuntos, $E_{G(0)}$ e $E_{G(1)}$, que contém, respectivamente, os pares ordenados com os *tokens* w_g e pesos p_g para a classe 0 e 1 para $\mathsf{F}_{(g)}$ treinado em D_{train}. Para maior clareza, a partir deste ponto sempre que mencionarmos $E_{L(n)}$ e $E_{G(n)}$, estaremos nos referindo a [$E_{L(0)}, E_{L(1)}$] e [$E_{G(0)}, E_{G(1)}$], respectivamente.

4.6 Extrair padrões de investigação

No escopo deste trabalho, os padrões de investigação correspondem a combinações de n palavras recorrentes nas sentenças de erro. Tais padrões são extraídos a partir de interseções frequentes com sentenças semelhantes, conforme a Equação 4.8, e considerados relevantes pelos modelos segundo a Equação 4.7. A hipótese central é que esses padrões, por estarem frequentemente presentes em sentenças semelhantes (mesma/oposta) possam funcionar como artefatos ou confundidores, induzindo o modelo a erros na classificação. Assim, padrões que ocorrem frequentemente em sentenças semelhantes da mesma classe de erro podem ser utilizados como atalhos de decisão, levando à dependência do modelo. Por outro lado, quando padrões relevantes para uma classe são encontrados em sentenças da classe oposta, isso pode indicar generalização inadequada, evidenciando correlações espúrias.

Nessa etapa, propomos a quantificação da frequência dos padrões em ambas as classes como uma variável útil para medir o grau de espuriedade desses padrões. Um padrão que ocorre com frequência nas duas classes, especialmente em contextos onde deveria ser distintivo, pode ser considerado potencialmente espúrio e, portanto, merece atenção especial na interpretação dos resultados do modelo. Definimos padrão de investigação P como combinações de n palavras que satisfazem aos seguintes critérios: (i) aparecem em sentenças de erro de uma classe específica e em sentenças semelhantes da classe oposta; (ii) são compostos pelas m palavras mais importantes para a classe; e (iii) estão presentes tanto na sentença de erro quanto nas sentenças semelhantes da mesma classe e da classe oposta. A definição formal é dada por:

$$P = \{p \mid p = (w_1, \ldots, w_n), w_i \in \mathsf{Top}_m(E_{\mathsf{G}(y)}), p \subseteq s_e, p \in \mathsf{M}(s_e) \wedge p \in \mathsf{O}(s_e)\} \quad (4.8)$$

onde p representa uma combinação de palavras $(w_1, w_2, \ldots, w_n)$, $w_i \in \mathsf{Top}_m(E_{\mathsf{G}(y)})$ indica que cada palavra do padrão está entre as m mais importantes para a classe y, $p \subseteq s_e$ significa que a sentença de erro s_e contém o padrão, e $p \in \mathsf{M}(s_e) \wedge p \in \mathsf{O}(s_e)$ indica que p aparece em sentenças semelhantes tanto da mesma classe ($\mathsf{M}(s_e)$) quanto da classe oposta ($\mathsf{O}(s_e)$).

Elaboramos o Algoritmo 2 para quantificar as ocorrências dos padrões em sentenças semelhantes, intra e entre classes, utilizando uma abordagem baseada na frequência de ocorrência em contextos opostos.

Algorithm 2 Algoritmo para reconhecimento dos padrões de investigação.

Require: D_{erro}, R.
Ensure: R_{freq}: Estrutura contendo frequências dos padrões para a mesma classe e a classe oposta, organizada por classe de erro.

1: $R_{\text{freq}} \leftarrow \{0 : \{\text{mesma} : \{\}, \text{oposta} : \{\}\}, 1 : \{\text{mesma} : \{\}, \text{oposta} : \{\}\}\}$
2: **for** cada classe $y \in \{0, 1\}$ **do**
3: **for** cada $s_e \in D_{\text{erro}}$ **where** classe$(s_e) = y$ **do**
4: $Palavras\_Importantes \leftarrow$ recuperar_palavras_importantes$(s_e, 1 - y)$
5: **for** $i \in \{0, 1\}$ **do**
6: $S_{sim} \leftarrow R[y][s_e][i]$ Sentenças similares para s_e na mesma ou oposta
7: $Combinacoes \leftarrow$ gerar_combinacoes$(Palavras_Importantes, S_{sim})$
8: **for** $p \in Combinacoes$ **do**
9: $contagem \leftarrow$ conta_ocorrencias(p, S_{sim})
10: **if** $i = 0$ **then**
11: $R_{\text{freq}}[y][\text{mesma}][p] \leftarrow R_{\text{freq}}[y][\text{mesma}].get(p, 0) + contagem$
12: **else**
13: $R_{\text{freq}}[y][\text{oposta}][p] \leftarrow R_{\text{freq}}[y][\text{oposta}].get(p, 0) + contagem$
14: **end if**
15: **end for**
16: **end for**
17: **end for**
18: **end for**
19: **return** R_{freq}

O Algoritmo inicia definindo uma estrutura R_{freq} para armazenar os resultados da quantificação dos padrões, organizada por classe e tipo (passo 1). Essa estrutura é projetada para separar os padrões encontrados nas sentenças de erro e nas sentenças semelhantes da mesma classe e da classe oposta. Para cada classe y e para cada sentença de erro s_e com rótulo em y, (passos 2 e 3), o algoritmo procede com os seguintes passos:

1. **Recuperação de Palavras Importantes:** Para cada s_e, o algoritmo recupera as palavras importantes da classe oposta ao seu rótulo. O objetivo é identificar termos que, embora relevantes para a classe oposta,

ocorrem na mesma classe de s_e — inclusive na própria s_e — e podem estar contribuindo para a decisão incorreta do modelo.

2. **Análise de Sentenças Semelhantes:** Para cada $i \in \{0, 1\}$, o algoritmo então recupera as sentenças semelhantes armazenadas em R (passo 6), específicas para cada sentença de erro s_e e classificadas por similaridade dentro da mesma classe ou entre classes opostas, definidas por i. (0: mesma classe; 1: classe oposta), conforme Expressão 4.3, isso permite uma comparação direta de como os mesmos padrões ocorrem em diferentes contextos de classificação.

   Geração de Combinações: Para cada palavra importante selecionada em $Palavras\_Importantes$ (passo 4), a função $gerar\_combinacoes$ gera todas as combinações possíveis de 2 até n palavras. Cada combinação (e palavra individual) é mantida apenas se todas as palavras que a compõem ocorrem simultaneamente em pelo menos uma sentença semelhante da mesma classe de s_e e em pelo menos uma da classe oposta, independentemente da ordem em que aparecem. Por exemplo, na s_e (classe 0) “aquisição de materiais em caráter de urgência visando ações de combate ao coronavirus”. As palavras importantes (para a classe 1) extraídas foram “aquisição”, “materiais” e “combate”. A combinação “aquisição materiais” é gerada e mantida, pois ocorre simultaneamente em pelo menos uma sentença semelhante da mesma classe — “aquisição de materiais de limpeza em caráter urgência visando ações de combate ao coronavirus” (classe 0) — e em outra da classe oposta — “aquisição de materiais hospitalares para o hospital ... com a finalidade em atender pacientes infectados pelo covid 19” (classe 1). Isso indica que a combinação reúne palavras importantes à classe 1, mas aparece em sentenças de ambas as classes, sugerindo potencial espuriedade.

3. **Contagem e Registro de Frequências:** Para cada combinação, contabilizamos sua frequência nas sentenças similares na mesma e na classe oposta. As contagens são acumuladas e armazenadas em R_{freq} (passos 10 a 14).

Após a execução do algoritmo, a estrutura R_{freq} é preenchida com as frequências dos padrões p em sentenças semelhantes e em sentenças de erro, considerando tanto a mesma classe y quanto a classe oposta. A frequência do padrão p em sentenças semelhantes da mesma classe é definida como a métrica a seguir: essa função, representada pela métrica μ_t, indica o número de vezes que o padrão p ocorre em sentenças semelhantes rotuladas na mesma classe y. De forma análoga, a frequência do padrão p em sentenças semelhantes da classe oposta é dada por:

$$\phi_t = f_{\text{train}}(y, \text{oposta}, p) = R_{\text{freq}}[y][\text{oposta}][p] \tag{4.9}$$

esse valor, denominado ϕ_t, expressa a frequência com que o padrão p ocorre em sentenças semelhantes rotuladas na classe oposta a y. Além disso, contabilizamos as frequências dos padrões nas sentenças de erro. A frequência do padrão p em sentenças de erro rotuladas na mesma classe y é calculada por:

$$\mu_e = f_{\text{error}}(y, \text{mesma}, p) = \sum_{s_e \in D_{\text{erro}}} \mathbb{I}(p \subseteq s_e \wedge \text{classe}(s_e) = y) \tag{4.10}$$

essa contagem corresponde à métrica μ_e, indicando quantas vezes o padrão ocorre em sentenças de erro cuja classe é y. Por fim, a frequência do padrão p em sentenças de erro da classe oposta é definida por:

$$\phi_e = f_{\text{error}}(y, \text{oposta}, p) = \sum_{s_e \in D_{\text{erro}}} \mathbb{I}(p \subseteq s_e \wedge \text{classe}(s_e) \neq y) \tag{4.11}$$

esta métrica, representada por ϕ_e, corresponde à ocorrência do padrão p em sentenças de erro rotuladas na classe oposta à y.

Para facilitar a análise dos padrões, definimos as razões κ_t e κ_e, que expressam o equilíbrio de ocorrência do padrão em diferentes contextos:

- κ_t: razão entre a frequência em sentenças semelhantes da classe oposta e da mesma classe:

  $$\kappa_t = \frac{\phi_t}{\mu_t} \tag{4.12}$$

  valores superiores a 1 sugerem que o padrão é mais recorrente na classe oposta, podendo indicar uma possível fonte de confusão para o modelo.

- κ_e: razão entre a frequência em sentenças de erro da classe oposta e da mesma classe:

$$\kappa_e = \frac{\phi_e}{\mu_e} \tag{4.13}$$

  um valor próximo de zero pode indicar que o padrão está fortemente associado a erros de classificação, principalmente ao favorecer incorretamente a classe oposta.

As variáveis calculadas nesta etapa serão consolidadas com as obtidas na etapa de identificação de padrões espúrios (Seção 4.7), e posteriormente utilizadas no processo de clusterização dos padrões (Seção 4.8).

4.7 Extrair potenciais padrões espúrios

Essa etapa tem como objetivo identificar potenciais padrões espúrios por meio da análise das variações nas métricas associadas ao desempenho do modelo quando esses padrões são removidos de D_{test}. A ideia é avaliar o impacto desses padrões no modelo, verificando se sua presença influencia diretamente as decisões do modelo em relação a uma classe específica. Isso permite observar mudanças nas métricas e, consequentemente, identificar acertos ou erros de previsão. A análise é generalizada para a a base de dados por meio da aplicação de validação cruzada. O Algoritmo 3 descreve os passos utilizados para essa seleção.

Dentre os parâmetros do algoritmo, m é o único ainda não abordado, ele corresponde ao conjunto de métricas originais calculadas com base no desempenho do modelo $\mathsf{F}_{(g)}$ sobre a base de dados D_{test}. Este conjunto inclui métricas especificidade (esp), sensibilidade (sen), verdadeiros positivos (fp) e verdadeiros negativos (fn). Essas métricas são empregadas como referências para avaliar as alterações decorrentes da perturbação dos dados. O algoritmo começa inicializando uma estrutura R_{pot} (passo 1) que armazenará os padrões e as métricas associadas a perturbação da base de teste. Para cada classe y e para cada padrão p associados a essa classe (passos 2 e 3), o algoritmo cria a versão perturbada D'_{test} removendo p de D_{test} (passo 4).

Algorithm 3 Seleção de candidatos a padrões espúrios

Require: D_{test}, $\mathsf{F}_{(g)}$, m, P, Y
Ensure: R_{pot}
1: $R_{\text{pot}} \leftarrow \{\}$
2: **for** $y \in Y$ **do**
3: **for** $p \in P$ **do**
4: $D'_{test} \leftarrow$ perturbar_base(D_{test}, p)
5: $m' \leftarrow$ avaliar_modelo$(\mathsf{F}_{(g)}, D'_{test})$
6: $\delta \leftarrow$ calcular_delta(m, m')
7: **if** $(\delta.esp > 0) \vee (\delta.sen > 0)$ **then**
8: $(\varrho_p, \tau_p, \varrho_a, \tau_a) \leftarrow$ calcular_metricas_perturbacao$(\delta.esp, \delta.sen, \delta.tp, \delta.tn)$
9: $R_{\text{pot}}[y]$.append$(p, (\varrho_p, \tau_p, \varrho_a, \tau_a))$
10: **end if**
11: **end for**
12: **end for**
13: **return** R_{pot}

Cálculo das Métricas: O modelo $\mathsf{F}_{(g)}$ é aplicado ao conjunto de teste perturbado, gerando um novo conjunto de métricas m' (passo 5). Em seguida, calcula-se a diferença entre as métricas originais m e as modificadas m' (passo 6). Diferenças positivas em especificidade ou sensibilidade indicam que a remoção do padrão p melhora o desempenho, sugerindo uma possível correlação espúria (Soares et al. 2024). Nesses casos, o padrão e as métricas associadas são adicionados à R_{pot} (passos 7-9), acompanhadas das seguintes métricas adicionais:

- **Quantidade de Sentenças Perturbadas (ϱ_p)**:
  - Soma dos valores absolutos das variações nos verdadeiros positivos e verdadeiros negativos ($|\delta.tp| + |\delta.tn|$). Reflete o total de alterações nas classificações, indicando uma medida do impacto da perturbação em termos da quantidade de sentenças cujos resultados foram afetados.

- **Quantidade de Acertos (ϱ_a)**:
  - Quantidade de sentenças que o modelo passou a acertar. É a soma das variações dos verdadeiros positivos e verdadeiros negativos ($\delta.tp+\delta.tn$), refletindo o efeito líquido dos acertos.
- **Taxa de Perturbação (τ_p)**:
  - Soma dos valores absolutos das variações da especificidade e sensibilidade ($|\delta.esp| + |\delta.sen|$). Representa a soma das magnitudes das variações da especificidade e sensibilidade. A soma dessas variações é útil para avaliar o impacto geral da perturbação, pois mesmo que uma métrica melhore e a outra piore, a soma das magnitudes dessas variações nos dá um entendimento do nível de perturbação que o modelo experimentou como um todo.
- **Taxa de Acerto (τ_a)**:
  - Soma das diferenças da sensibilidade e especificidade ($\delta.esp + \delta.sen$). Esta soma fornece uma visão global de como as perturbações afetaram ambas as métricas em conjunto, refletindo o impacto líquido dessa taxa no modelo.

O resultado do algoritmo é armazenado na variável R_{pot}, uma estrutura bidimensional, com uma dimensão para cada classe y. Cada dimensão contém uma lista de dicionários, onde as chaves são os padrões e os valores são as métricas associadas a eles (passo 9). Este resultado será unificado ao resultado da etapa “Extrair padrões de investigação” para o cálculo do grau de espuriedade de cada padrão, como veremos na próxima Seção.

4.8 Clusterizar padrões

Nesta seção, abordamos a etapa de “Clusterização de padrões”, em que utilizamos técnicas de aprendizado não supervisionado para formar grupos lógicos e interpretáveis. O critério de interpretação dos *clusters* adotado neste trabalho baseia-se na hipótese de que padrões mais distantes dos

centroides tendem a indicar associações espúrias ou ruído, conforme estudos que relacionam a distância ao centroide à detecção de *outliers* e à presença de dados anômalos em algoritmos de *k-means*e variantes (Gan e Ng 2017, Gan 2025). Em contextos mais complexos, métricas robustas de distância, como o *Kernelized Rank-Order Distance,* têm sido propostas para reduzir o impacto de ruído e preservar a estrutura dos dados (Huang, Wang e Zhu 2020).

A etapa é implementada por meio do Algoritmo 4, que define um conjunto de passos, os quais podem ser executados utilizando diferentes técnicas de clusterização. Como entrada, o algoritmo utiliza as métricas obtidas nas Seções 4.6 e 4.7, concatenadas com base nos padrões comuns a R_{freq} e R_{pot}, identificados por cada padrão p e sua classe y. A Tabela 5 apresenta as variáveis utilizadas na clusterização, incluindo p e y, que orientam a formação dos grupos.

Tabela 5 – Variáveis de Entrada para a Clusterização de Padrões, onde p é um padrão comum a R_{freq} e R_{pot}.

Métrica	**Descrição**
p	Padrão selecionado com potencial padrão espúrio.
y	Classe cujo padrão é considerado importante.
ρ	Peso global do padrão.
μ_t	Frequência do padrão em sentenças semelhantes na mesma classe.
ϕ_t	Frequência do padrão em sentenças semelhantes na classe oposta.
μ_e	Frequência do padrão em sentenças de erros na mesma classe.
ϕ_e	Frequência do padrão em sentenças de erros na classe oposta.
τ_p	Taxa de perturbação.
ϱ_p	Quantidade de perturbação observada.
τ_a	Taxa de acerto após perturbação.
ϱ_a	Quantidade de acertos corrigidos.
$\kappa_t = \frac{\phi_t}{\mu_t}$	Razão da frequência do padrão nas sentenças semelhantes da classe oposta (ϕ_t) e na mesma classe (μ_t).
$\kappa_e = \frac{\phi_e}{\mu_e}$	Razão da frequência do padrão nas sentenças de erro da classe oposta (ϕ_e) e na mesma classe (μ_e).

O Algoritmo 4 detalha o processo de clusterização e visualização dos padrões com base nas métricas da Tabela 5. Ele organiza o agrupamento dos padrões, permitindo explorar suas similaridades, diferenças e estimar o grau de espuriedade a partir de métricas quantitativas.

Inicialmente, as métricas de entrada são normalizadas (passo 1) para assegurar comparabilidade entre variáveis, reduzindo possíveis distorções

Algorithm 4 Clusterização de Padrões

Require: Métricas da Tabela 5.
Ensure: *Clusters* de padrões
1: Normalizar as métricas de entrada
2: Reduzir dimensionalidade.
3: Calcular o número ótimo de *clusters*.
4: Aplicar o algoritmo e clusterização.
5: Gerar gráficos para interpretação.
6: **return**

associadas a escalas diferentes. Posteriormente, uma técnica de redução de dimensionalidade é aplicada (passo 2), com o objetivo de preservar as informações mais relevantes, simplificando a análise e facilitando a visualização dos agrupamentos. O número ideal de *clusters* é estimado por meio de métodos baseados na avaliação da qualidade da segmentação (passo 3), buscando um equilíbrio entre coesão interna e separação entre grupos. Para a clusterização, emprega-se o algoritmo *K-Means* (passo 4), que permite calcular a distância dos padrões ao centroide, adotada como hipótese de espuriedade neste trabalho.

Alternativamente, técnicas como *DBSCAN* (Ester et al. 1996) ou *HDBSCAN* (Campello, Moulavi e Sander 2013) poderiam ser utilizadas na etapa de agrupamento. Nesses casos, o grau de espuriedade pode ser estimado pela densidade local dos padrões ou pela distância em relação aos pontos centrais de densidade ou membros centrais dos grupos. Embora essas abordagens não utilizem centroides explícitos, a densidade ou a posição relativa dos padrões no grupo podem fornecer indícios de isolamento ou atipicidade, alinhando-se ao objetivo de detectar padrões com maior potencial de espuriedade.

Essa abordagem permite que as variáveis sejam analisadas de forma mais ampla, sem depender de suposições prévias sobre quais padrões são espúrios. Dessa forma, o algoritmo pode descobrir associações inesperadas entre os padrões, o que não seria facilmente perceptível nas etapas anteriores. No próximo capítulo, são descritos os experimentos realizados e os resultados obtidos na aplicação do método às bases Contratos, Licitações e IMDB, permitindo avaliar sua eficácia e compará-lo a abordagens de referência.

5 Resultados e Discussões

Este capítulo apresenta os experimentos e resultados obtidos com a aplicação do método proposto. Inicialmente, descrevemos a configuração experimental adotada, na sequência, apresentamos os resultados obtidos nas bases de dados Contratos (C) e Licitações (L). Por fim, reportamos os experimentos realizados na base IMDB, com o objetivo de analisar a capacidade de generalização do método e compará-lo com abordagens de referência na literatura. Os experimentos foram executados em ambiente Python 3.9, com as bibliotecas `Scikit-Learn 1.4.1`, `Torch 2.1.1` e `Transformers 4.35.2`, utilizando os modelos descritos na Seção 2.1.1.

5.1 Configuração Experimental

Esta seção descreve os procedimentos adotados para implementar e testar o método proposto. Detalhamos as técnicas aplicadas em cada etapa e os respectivos parâmetros utilizados.

Pré-processamento: Adotamos fluxos de pré-processamento de acordo com a representação textual utilizada por cada modelo. Para os modelos lineares, foram aplicadas duas estratégias de representação: *TF-IDF* e WE. No caso do *TF-IDF*, o texto foi submetido às seguintes etapas: limpeza (remoção de pontuações e caracteres especiais), tokenização, normalização (remoção de acentuação e conversão para minúsculas) e remoção de *stop words*. Para WE, as sentenças foram processadas com o modelo pré-treinado BERTimbau (Souza, Nogueira e Lotufo 2020), voltado para a língua portuguesa, utilizando a arquitetura *Sentence-BERT (SBERT)* (Reimers e Gurevych 2019) para gerar *embeddings* vetoriais densos e fixos.

Para o modelo BERTimbau, o texto foi tokenizado em *subwords* por meio do tokenizador *WordPiece*, especificamente o disponibilizado com o modelo. Foram adicionados os tokens especiais [CLS] e [SEP], conforme a arquitetura proposta por (Devlin et al. 2018). O parâmetro *max_length* foi configurado para

512, com a opção de truncamento ativada, assegurando que textos superiores a esse limite fossem cortados. O preenchimento (*padding*) foi ajustado para *True*, garantindo que todas as sequências tivessem comprimento uniforme. A entrada final do modelo é composta pelos índices dos *subwords*, sendo que os *segment embeddings* e *positional embeddings* são incorporados internamente durante o processamento.

Treinar/Testar Modelos: Neste experimento, utilizamos dois modelos lineares (LRG e SVM) e o modelo pré-treinado BERTimbau. Os modelos LRG e SVM foram avaliados com vetores gerados a partir de TF-IDF e de WE, conforme descrito anteriormente. Nesses casos, as representações textuais são pré-computadas e permanecem inalteradas durante o treinamento. No processo de treinamento, empregamos validação cruzada com $k = 5$ e otimização dos hiperparâmetros via `BayesSearchCV`, ajustando C no LRG e α no SVM. O modelo LRG utiliza regularização L2 com até 1000 iterações. O SVM aplica o otimizador *Stochastic Gradient Descent (SGD)* para atualização eficiente dos parâmetros. Por outro lado, o modelo BERTimbau, utilizado como classificador, realiza ajuste fino de seus *embeddings* contextuais ao longo do processo de treinamento. As representações textuais são ajustadas por meio de *fine-tuning*, permitindo a otimização dos *embeddings* ao longo do treinamento. Utilizamos sua arquitetura pré-treinada, complementada por uma camada densa final com ativação *softmax* para a tarefa de classificação binária. O treinamento foi realizado com o otimizador Adam, taxa de aprendizado de 2×10^{-5}, por três épocas, com tamanho de lote igual a 4. Utilizamos *dropout* (0,1), *early stopping* e *checkpointing* para evitar sobreajuste. A Tabela 6 apresenta as configurações dos modelos.

Analisar erros do modelo: Para identificar as sentenças semelhantes, utilizamos estratégias distintas conforme a representação textual adotada. Nos modelos lineares (LRG e SVM), as sentenças foram vetorizadas com TF-IDF ou WE, e a similaridade foi calculada por cosseno (Huang, Yin e Hou 2011). Já no modelo BERT, aplicamos o BERTScore (Zhang et al. 2019), que utiliza *embeddings* contextualizados para aferir a similaridade semântica entre as sentenças. Em ambos os casos, adotamos valores iniciais de $\lambda_m = \lambda_o = 0{,}9$,

Tabela 6 – Configuração dos modelos. RT: Representação Textual; OH: Otimização de Hiperparâmetros; VC: Validação Cruzada; OT: Otimizador; TA: Taxa de Aprendizado; REG: Regularização; TL: Tamanho do Lote; EP: Épocas.

Configuração	**LRG**	**SVM**
RT	TF / WE	TF-IDF / WE
OH	C	α
VC	$k = 5$	$k = 5$
OT	–	SGD

(a) Modelos lineares

Configuração	**BERT**
RT	*Embeddings*
TL	4
EP	3
OT	Adam
TA	2×10^{-5}
REG	Drop = 0,1; ES

(b) Modelo BERT

com decrementos $\delta_m = \delta_o = 0{,}05$ por iteração, conforme o Algoritmo 1.

Extrair palavras importantes: Para esta etapa, adotamos o método de seleção de palavras importantes proposto por (Soares et al. 2024), que utiliza o LIME como explicador local. A abordagem baseada no LIME apresentou resultados promissores em bases semelhantes às utilizadas neste estudo, especialmente para os modelos SVM e BERTimbau. Para o modelo LRG, utilizamos seu próprio mecanismo interno de explicação, baseado nos coeficientes gerados durante o treinamento. Esses coeficientes indicam a magnitude e a direção do efeito de cada palavra nas previsões do modelo. Conforme a Equação 4.6, o explicador $E_{xp}(\mathsf{F}_{(g)}, s)$ associa a cada sentença $s \in S$ um conjunto $E_{L(n)}$ de pares (w_l, p_l), onde w_l é uma palavra e p_l seu peso para a classe n. Em explicadores intrínsecos, E_{xp} representa esse mecanismo interno, como o cálculo de pesos associados às variáveis independentes, que quantificam sua influência sobre a variável resposta (Ali et al. 2023). Em todos os casos, a explicação global $E_{G(n)}$, definida na Equação 4.7, foi obtida calculando a raiz quadrada da soma dos pesos das palavra em cada classe, conforme descrito em (Ribeiro, Singh e Guestrin 2016). A explicação global define as ordem das palavras mais importantes por classe, que é utilizada na etapa seguinte do método proposto.

Extrair padrões de investigação: Nesta etapa, utilizamos os modelos treinados em cada *fold* da validação cruzada e aplicamos perturbações à base

de teste com base nas ($m = 60$) palavras mais importantes de cada classe, previamente obtidas na Etapa “Extrair palavras importantes”. Essas palavras correspondem aos *tokens* com maiores pesos p_g nas explicações globais $E_{G(n)}$, calculadas a partir dos explicadores adotados para cada modelo — coeficientes internos da LRG, pesos do LIME aplicados ao SVM e ao BERT.

Os padrões extraídos consistem em combinações de até três palavras ($n = 3$), geradas segundo os critérios definidos na Equação 4.8. A identificação e quantificação desses padrões foram realizadas conforme o Algoritmo 2, o qual verifica a ocorrência das combinações em sentenças de erro e em sentenças semanticamente semelhantes, tanto da mesma classe quanto da classe oposta.

Clusterizar padrões: Esta etapa aplica o Algoritmo 4 para agrupar padrões com base nas métricas definidas na Tabela 5, formando estruturas logicamente interpretáveis. A seguir, são descritas as técnicas utilizadas em cada etapa do algoritmo neste experimento:

1. **Normalização**: As métricas são padronizadas com *Standard Scaler*, assegurando média zero e desvio padrão unitário, o que evita distorções causadas por diferenças de escala entre variáveis.

2. **Redução de Dimensionalidade**: Utilizamos a técnica *Principal Component Analysis* (PCA) para projetar os dados componentes principais, reduzindo a complexidade com preservação da maior parte da variância explicativa.

3. **Número de Clusters**: O número ótimo de *clusters* é estimado pelo método do cotovelo, com base na *Within-Cluster Sum of Squares* (WCSS).

4. **Clusterização**: O algoritmo *K-Means* é utilizado por alinhar-se à hipótese central deste trabalho, segundo a qual padrões mais distantes de seus centroides indicam maior potencial de espuriedade. O método agrupa os dados minimizando a variância intra-cluster, favorecendo interpretações baseadas em distância, conforme discutido em (He, Xu e Deng 2003, Gan e Ng 2017, Huang, Wang e Zhu 2020, Gan 2025).

5. **Visualização**: As distribuições dos *clusters* são visualizadas em 1D, 2D e 3D com base nos componentes principais, facilitando a análise e a identificação de agrupamentos e *outliers*.

5.2 Resultados

Conduzimos os experimentos considerando 20 combinações, resultantes da variação entre modelos de classificação (SVM, Regressão Logística e BERTimbau), representações textuais (TF-IDF e WE), conjuntos de dados (Contratos e Licitações) e classes-alvo (0 e 1). SVM e Regressão Logística foram avaliados com ambas as representações, enquanto o BERTimbau foi testado apenas com WE. Embora todas as combinações tenham seguido as configurações descritas na Seção 5.1, optamos por detalhar apenas um cenário. Essa escolha atende a dois propósitos: (*i*) apresentar todas as combinações tornaria o documento desnecessariamente extenso e redundante, comprometendo a clareza; e (*ii*) o cenário selecionado representa de forma adequada as principais tendências observadas e ilustra as hipóteses e objetivos do estudo.

As etapas de pré-processamento e de treino/teste revelaram relações diretas entre as características dos conjuntos de dados, a representação textual e o desempenho dos modelos. A Tabela 7 resume os resultados obtidos.

Na base Contratos, que apresenta menor diversidade lexical (5.662 palavras distintas) e sentenças mais curtas (média de 30,5 palavras), os modelos lineares com TF-IDF obtiveram os melhores resultados: o LRG alcançou 94,58% de acurácia, 94,24% de sensibilidade e 94,93% de especificidade; o SVM apresentou valores próximos, com 94,20% de acurácia, 94,34% de sensibilidade e 94,90% de especificidade. Já com WE, houve queda no desempenho: LRG obteve 84,23% de acurácia e SVM, 86,03%. O modelo BERT, que utiliza *embeddings* contextuais ajustáveis, manteve desempenho elevado, com 94,97% de acurácia, 95,39% de sensibilidade e 94,93% de especificidade.

A base Licitações, por sua vez, apresentou maior diversidade lexical (7.960 palavras distintas) e sentenças mais longas (média de 43,95 palavras).

Tabela 7 – Desempenho dos modelos nas bases de Contratos e Licitações com diferentes representações textuais. Abreviações: **TR** = Representação Textual, **ACC** = Acurácia, **SEN** = Sensibilidade, **ESP** = Especificidade.

Base de dados	Modelo	RT	ACC (%)	SEN (%)	ESP (%)
Contratos	LRG	TF-IDF	94.58	94.24	94.93
		WE	84.23	83.19	85.27
	SVM	TF-IDF	94.20	94.34	94.90
		WE	86.03	81.48	90.58
	BERT	Embeddings	**94.97**	**95.39**	**94.93**
Licitações	LRG	TF-IDF	**97.92**	**97.88**	**97.96**
		WE	93.16	93.71	92.63
	SVM	TF-IDF	97.52	97.52	97.54
		WE	93.80	94.36	93.26
	BERT	Embeddings	97.58	**97.61**	**97.55**

Os modelos lineares tiveram melhor desempenho com TF-IDF: LRG obteve 97,92% de acurácia, enquanto o SVM atingiu 97,52%. Com WE, as taxas caíram para 93,16% e 93,80%, respectivamente. O modelo BERT, por sua natureza contextual, apresentou desempenho estável: 97,58% de acurácia, 97,61% de sensibilidade e 97,55% de especificidade.

O BERT lida de forma eficaz com dados que apresentam alta diversidade lexical e estruturas textuais complexas, mantendo um desempenho consistente em ambos os conjuntos de dados. Já os modelos LRG e SVM apresentam bons resultados com TF-IDF em conjuntos com vocabulário mais restrito e sentenças mais curtas, mas têm queda de desempenho ao utilizar WE, demonstrando dificuldade em explorar plenamente suas informações semânticas mais ricas.

A análise dos erros dos modelos destacou padrões específicos que contribuíram para classificações incorretas. A Tabela 8 apresenta a sentença mais representativa, da classe Aquisições específicas da saúde (1), por modelo e tipo de representação textual na base de Contratos. Essas sentenças influenciaram erros na classe Outras aquisições (0). O método identifica a sentença representativa como aquela semelhante que mais frequentemente ocorre na classe oposta. Os modelos SVM e LRG com WE tenderam a confundir sentenças relacionadas a processos de aquisição. Em especial, o modelo SVM classificou incorretamente sentenças semelhantes a *, sugerindo que

a representação baseada em WE capturou relações semânticas entre diferentes categorias de aquisição, mas falhou em distingui-las corretamente.

Tabela 8 – Sentenças da classe 1 (aquisições específicas de saúde) mais frequentemente identificadas como semelhantes às sentenças de erro da classe 0 (outras aquisições) na base de dados de Contratos.

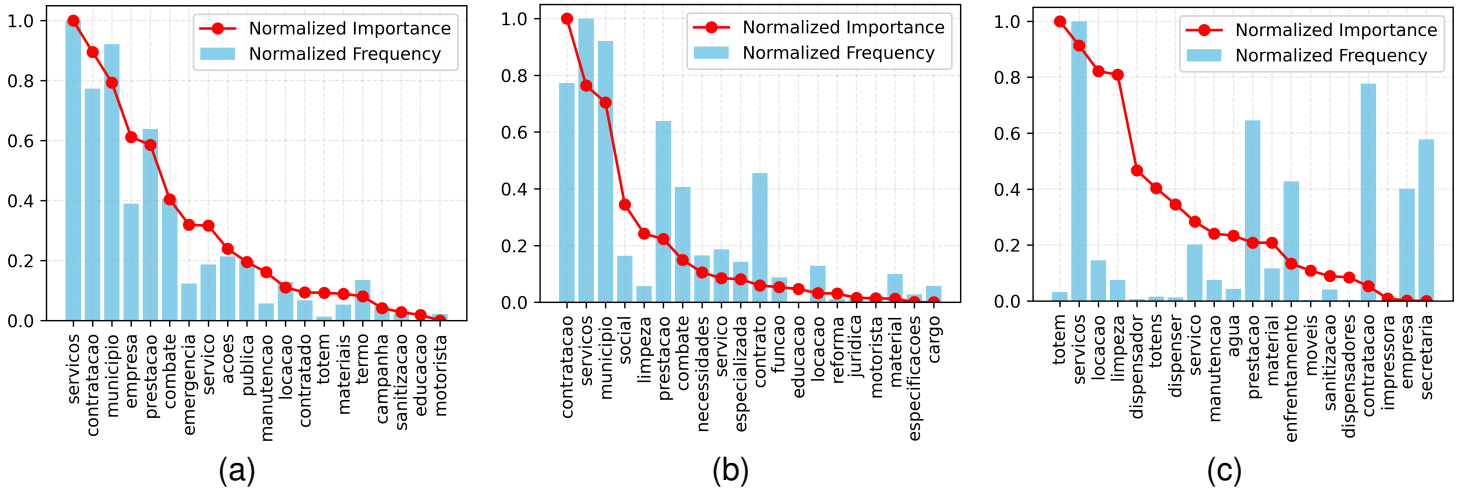
Modelo	**RT**	**Sentenças mais representativas**
SVM	WE	"Aquisição de insumos médicos hospitalares em caráter emergencial de acordo com a lei federal... para atender as necessidades da secretaria municipal de saúde de... no combate à pandemia do COVID-19."*
	TF-IDF	"Aquisição de testes rápidos destinados à secretaria municipal de saúde."‡
LRG	WE	"Aquisição de medicamentos para a secretaria municipal de saúde de..."§
	TF-IDF	"Contrato de aquisição de álcool gel para atender às necessidades da secretaria municipal de saúde de... para enfrentamento da emergência de saúde pública decorrente do coronavírus."¶
BERT	Embeddings	"Contratação de empresa para fornecimento de testes rápidos."†

O modelo BERT apresentou dificuldades com sentenças relacionadas à contratação de serviços. A sentença mais representativa † indica que o modelo pode confundir contratos semelhantes que envolvem diferentes tipos de produtos e serviços. Por outro lado, os modelos SVM e LRG com TF-IDF apresentaram erros associados a palavras frequentes. Por exemplo, o modelo LRG teve dificuldade com sentenças semelhantes a ¶, sugerindo que termos altamente recorrentes como "contrato", "aquisição" e "municipal" levaram a uma superestimação de sua importância.

A análise dos resultados da Etapa "Extrair de Palavras Importantes" revelou diferenças na atribuição de importância entre os modelos com representações textuais distintas. A Figura 3 ilustra como importância e frequência se correlacionam para as 20 palavras mais relevantes no SVM com TF-IDF e WE, com foco na classe "Outras aquisições" (0) da base de Contratos.

O SVM com TF-IDF (Figura 3a) priorizou palavras frequentes como "contratação" e "serviços", evidenciando uma dependência da frequência das palavras em detrimento de seu significado semântico. Com WE (Figura 3b), houve uma leve mudança, incorporando termos menos frequentes, mas

Figura 3 – Comparação das palavras mais importantes para a classificação na base Contratos, classe 0. (a) SVM-TFIDF-C0, (b) SVM-WE-C0 e (c) BERTIMBAU-C0. Modelos lineares (a, b) dependem mais de palavras frequentes, enquanto o BERTIMBAU (c) captura termos semanticamente relevantes.



contextualmente relevantes, como “social” e “limpeza”, que, apesar de menos recorrentes, mantêm relação com o contexto de compras públicas — ainda que o viés de frequência permanecesse dominante, tendência também observada no modelo LRG. Já o BERT, utilizando *embeddings* contextuais (Figura 3c), destacou palavras menos frequentes, porém semanticamente mais significativas, como “totem” e “dispensador”, itens frequentemente adquiridos durante a pandemia de COVID-19, mas que não são exclusivos da área da saúde.

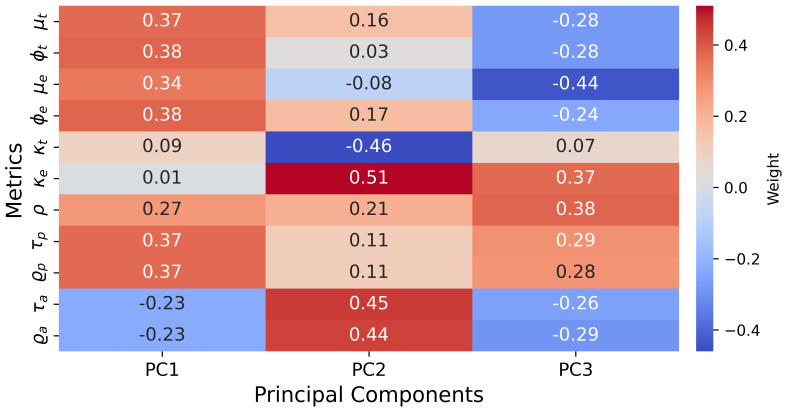
5.2.1 Clusterizar padrões

Os resultados das etapas “Extrair Padrões de Investigação” e “Extrair Padrões Espúrios Potenciais” não são apresentados separadamente, pois estão fortemente integrados à etapa “Clusterizar Padrões”. Apresentá-los de forma isolada traria redundância e prejudicaria a coesão dos achados. Assim, apresentamos os resultados do agrupamento de padrões aplicados ao modelo SVM com representação TF-IDF, no conjunto de dados Contratos, para a classe “Outras aquisições” (0). Esses resultados são comparados com os obtidos utilizando SVM com WE e BERT, e refletem a tendência geral observada.

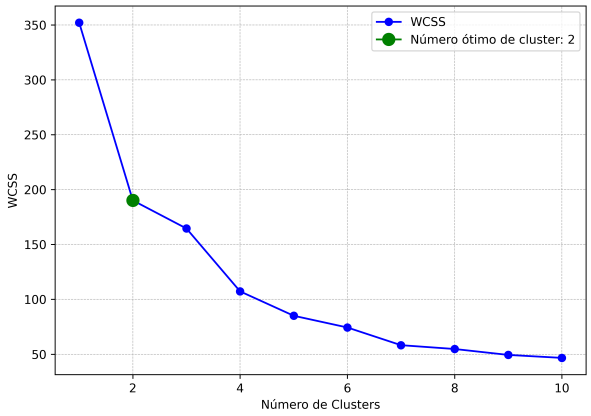
A Figura 4 apresenta dois aspectos complementares da análise dos padrões extraídos. A Figura 4a mostra o mapa de calor dos pesos atribuídos a cada métrica nos componentes principais gerados pelo PCA. Cores quentes (vermelho intenso) indicam contribuições positivas mais relevantes, enquanto cores frias (azul escuro) indicam contribuições negativas com maior influência. Valores próximos de zero aparecem em tons claros, sinalizando baixa contribuição. Já a Figura 4b exibe o resultado do método do cotovelo aplicado. O gráfico mostra a variação da soma dos quadrados intra-cluster (WCSS) em função do número de agrupamentos k. O ponto de inflexão, marcado no gráfico, indica que o número ótimo de *clusters* é $k = 2$.

Figura 4 – **SVM-TFIDF–C0**: Análise dos padrões da classe 0 gerados pelo modelo SVM com representação TF-IDF. A subfigura (a) mostra o mapa de calor dos pesos dos componentes principais (PCA), destacando a contribuição relativa de cada métrica. A subfigura (b) exibe o gráfico do método do cotovelo, cujo ponto de inflexão sugere $k = 2$ como número ótimo de agrupamentos.

(a) Mapa de calor dos pesos do PCA.



(b) Método do cotovelo.



5.2.2 Interpretação dos Agrupamentos na classe “Outras Aquisições”

Para o primeiro Componente Principal (PC1), as variáveis com maior contribuição são: ϕ_t (frequência em semelhantes da classe oposta); ϕ_e (frequência em erros da classe oposta), ambas com 0,38; μ_t (frequência em semelhantes da mesma classe); τ_p (taxa de perturbação); ϱ_p (quantidade de

perturbações), as três com 0,37; e μ_e (frequência em sentenças de erro da mesma classe), com 0,34. Esses valores indicam que o PC1 está fortemente relacionado às frequências de treino e erro, tanto na mesma classe quanto na classe oposta e às perturbações. O impacto é equilibrado entre variáveis que medem a frequência dos padrões e a quantidade de perturbações, sugerindo que PC1 captura as variações globais das frequências dos padrões no treino e teste com dados perturbados. Por outro lado, a variável κ_e (razão de frequência em sentenças de erro) tem a menor influência, com 0.01, indicando que essa relação entre frequências de erro não é relevante para este componente. Dessa forma, Podemos concluir que PC1 está relacionado ao comportamento do modelo frente às frequências de treino e erro e às perturbações, refletindo a sensibilidade a padrões muito frequentes e o impacto das perturbações no resultado final.

No PC2, as métricas mais influentes são: κ_t (razão de frequência em sentenças semelhantes), com $-0{,}46$; κ_e, com $0{,}51$; τ_a (taxa de acerto após a perturbação), com $0{,}45$; e ϱ_a (número de previsões corrigidas), com $0{,}44$. O PC2 captura o impacto relativo das correlações entre classes e os ganhos de desempenho obtidos com a remoção dos padrões analisados. A presença simultânea de κ_e e κ_t indica que o PC2 reflete a dependência do modelo em padrões frequentes em sentenças semelhantes e de erro na classe oposta, enquanto τ_a e ϱ_a evidenciam como essas relações podem ser exploradas para corrigir previsões incorretas. O equilíbrio entre as métricas de proporção e de correção destaca a interação entre a frequência relativa dos padrões e os ganhos de desempenho ajustado.

No PC3, as métricas relevantes são μ_e, com $-0{,}44$; ρ (peso global), com $0{,}38$; e κ_e, com $0{,}37$. Este componente principal foca na influência de padrões específicos com alta frequência em sentenças de erro e sua importância global. A presença simultânea de μ_e e ρ indica que o PC3 captura como padrões individualmente relevantes contribuem para os erros de classificação, refletindo um comportamento mais localizado no espaço dos erros.

Em resumo: **(i)** o PC1 reflete a influência das frequências dos padrões e das perturbações sobre as previsões; **(ii)** o PC2 representa as relações entre

frequências relativas e os ganhos de desempenho resultantes da remoção de padrões; e **(iii)** o PC3 enfatiza o impacto de padrões específicos nos erros de classificação. A Tabela 9 apresenta as contribuições das métricas para os componentes principais (PC1, PC2 e PC3) no modelo SVM-C0.

Tabela 9 – SVM-FTIDF-C0: contribuições das métricas para os componentes principais (PC1, PC2 e PC3).

	Métricas	**Valores**
PC1	μ_t, τ_p, ϱ_p	0,37
	ϕ_t, ϕ_e	0,38
	μ_e	0,34
PC2	κ_t	$-$0,46
	κ_e, τ_a, ϱ_a	0,51, 0,45, 0,44
PC3	μ_e	$-$0,44
	ρ, κ_e	0,38, 0,37

A análise dos *clusters* testou a hipótese de que padrões mais distantes dos centroides possuem maior potencial de serem espúrios, devido à sua variabilidade. Para validar essa hipótese, foram analisados os padrões dentro de cada agrupamento com base em suas projeções (distância ao centroide) nos componentes principais (PC1, PC2 e PC3) e nas respectivas métricas associadas. A Figura 5 apresenta uma visualização tridimensional do resultado do agrupamento obtido por meio do algoritmo K-means, com $k = 2$ (definido a partir do método do cotovelo para os três componentes).

5.2.2.1 Análise do Cluster 0

Padrão mais próximo: o padrão “empresa” apresenta distância de 0,192 em relação ao centroide. Esse padrão possui projeções equilibradas nos três componentes principais, alinhando-se às características médias do *cluster*. No PC1 (3,02), associada às métricas de frequência, o padrão apresenta valores moderados para μ_t e ϕ_t (frequência em semelhantes da classe oposta), indicando baixa frequência em sentenças semelhantes. No PC2 (0,95), que captura razões entre frequências, o padrão mostra valores consistentes para κ_t e κ_e, evidenciando estabilidade na relação representada pela métrica. Por fim, no PC3 (0,72), relacionada aos pesos globais e às frequências em sentenças

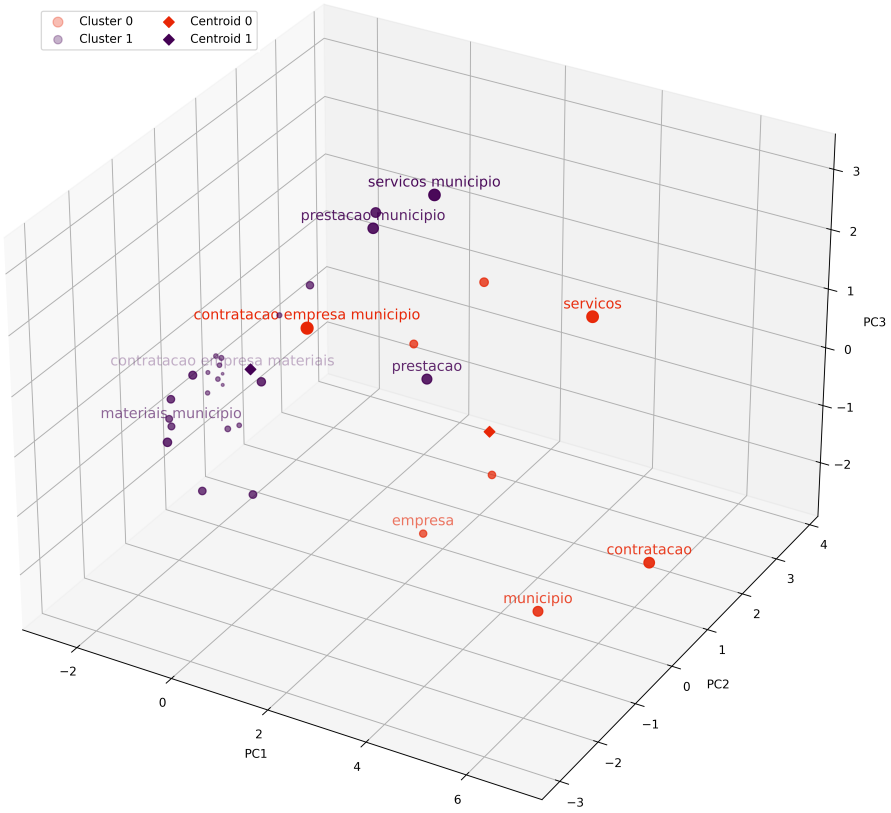


Figura 5 – **Visualização 3D do SVM-TFIDF-C0:** Representação dos agrupamentos em um espaço tridimensional. O *Cluster* 0 é representado em vermelho e o *Cluster* 1 em roxo, com os centroides indicados por losangos e os padrões por esferas. O tamanho e a opacidade das esferas são proporcionais à distância em relação aos respectivos centroides. Padrões mais distantes (esferas maiores e mais opacas) apresentam maior variabilidade e indicam maior potencial de correlação espúria.

de erro, o valor de ρ exerce influência moderada, indicando alinhamento com o comportamento médio do *cluster*.

Padrões mais distantes: Os padrões mais distantes do centroide apresentaram projeções extremas. O padrão “serviços”, com distância de 4,062, mostrou projeção elevada no PC1 (3,02) devido valores em ϕ_t (frequência em semelhantes da classe oposta) e ϕ_e (frequência em erros da classe oposta). No PC2 (3,76), esse padrão apresentou forte dependência das métricas de razão, especialmente κ_e, indicando influência significativa em contextos de erro. No PC3 (1,19), o peso global (ρ) reforçou a dependência do modelo em relação a esse padrão. O padrão “contratação empresa município”, com distância de

4,270, apresentou projeções elevadas no PC3 (3,16), impulsionadas por ρ e ϕ_e . Nas PC1 e PC2, os valores foram moderados, sugerindo que a influência desse padrão está concentrada principalmente em erros.

5.2.2.2 Análise do Cluster 1

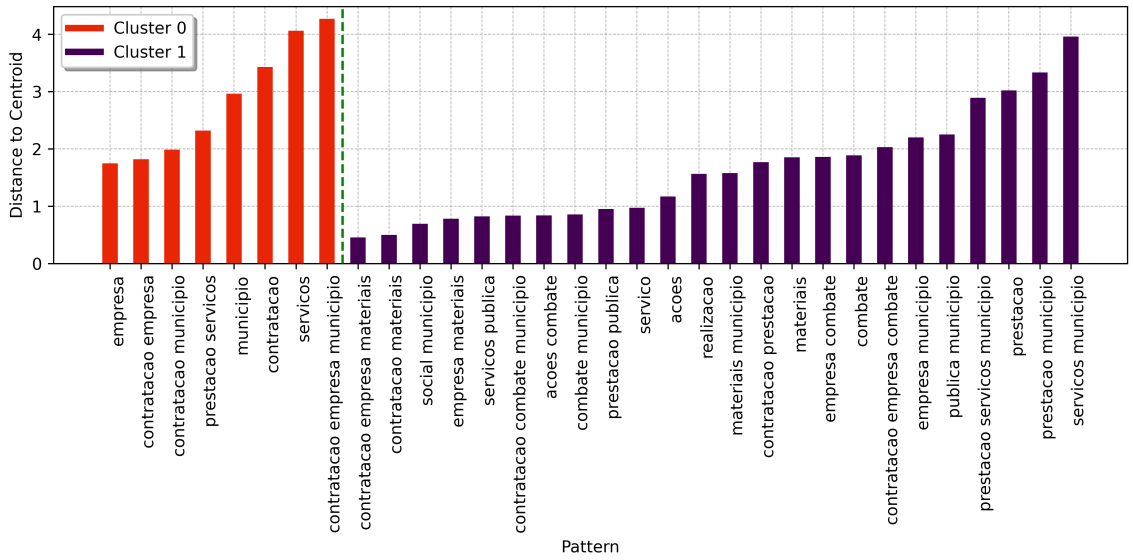
Padrão mais próximo: O padrão "contratacao empresa materiais" apresenta distância de 0,455 e valores típicos nos três componentes principais. No PC1 ($-$1,60), associada às métricas de frequência, o padrão reflete valores representativos para μ_t e ϕ_t (frequência em semelhantes da classe oposta). No PC2 ($-$0,28), relacionada às razões de frequência e correções, o padrão apresenta projeções compatíveis com κ_t e κ_e, indicando baixa variabilidade nesses contextos. Por fim, no PC3 ($-$0,03), que captura pesos globais e frequências em erros, o padrão é influenciado por ρ, reforçando seu alinhamento com as características médias do *cluster*.

Padrões mais distantes: Os padrões mais distantes, "prestação município", com distância de 3,33, e "serviços município", com distância de 3,96, apresentaram projeções extremas no *cluster*. O padrão "prestação município" se destacou no PC2 (3,33) devido à forte influência de κ_t e κ_e. Já o padrão "serviços município" apresentou valores elevados no PC2 (3,68), também influenciado por κ_e, e no PC3 (1,19), relacionada a ρ, destacando sua relevância nos erros de classificação e indicando maior variabilidade e potencial de correlação espúria.

5.2.2.3 Visualização das Distâncias aos centroides

Nesse trabalho, sugerimos que a distância dos padrões aos centroides pode ser utilizada como uma métrica adicional para avaliar o potencial espúrio do padrão. Padrões localizados mais distantes dos centroides tendem a ser menos representativos do agrupamento e, portanto, possuem maior probabilidade de corresponderem a correlações espúrias, como demonstrado nos exemplos anteriores. Essa distância reflete o grau de desvio de um padrão em relação ao comportamento típico do *cluster*, indicando sua inconsistência dentro do agrupamento (Gan e Ng 2017, Huang, Wang e Zhu 2020, Gan 2025).

Figura 6 – SVM-TFIDF-C0 – Gráfico da distância dos padrões aos centroides, da classe 0 (base Contratos), SVM com TF-IDF. *Cluster* 0 em vermelho e *Cluster* 1 em roxo.

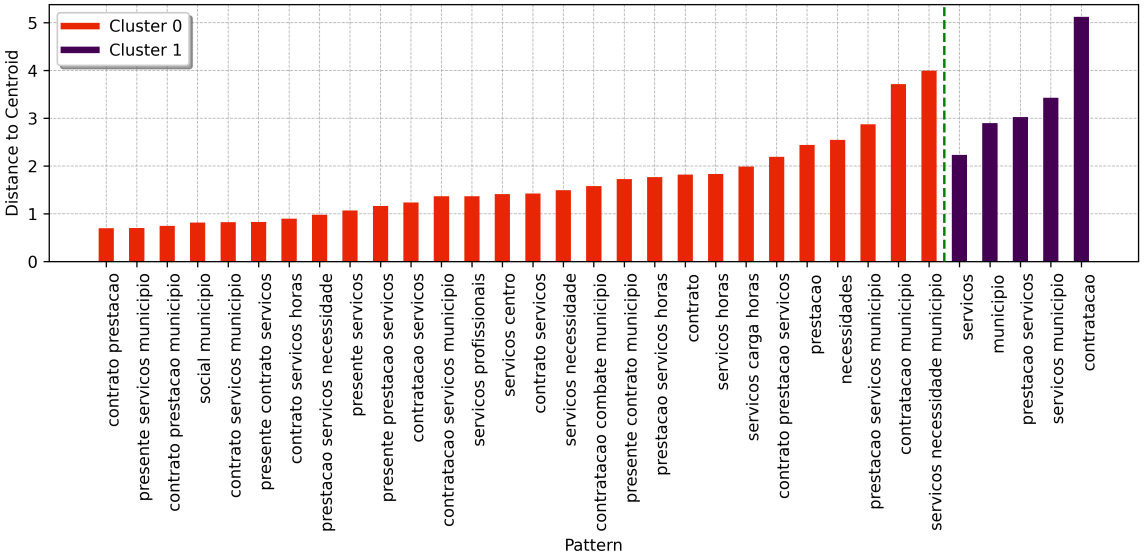


No modelo SVM com TF-IDF (Figura 6), os padrões mais próximos do centroide no *cluster* 0, “empresa” e “contratação empresa”, apresentam baixo peso global (ρ) e variações reduzidas nas razões de frequência κ_t e κ_e, indicando ocorrência estável. Por outro lado, os padrões mais distantes, “contratação empresa município” e “serviços”, possuem altos valores de ρ e discrepâncias significativas entre μ_t e ϕ_t, sugerindo que sua presença pode enviesar a classificação. No *cluster* 1, os padrões mais próximos, “contratação empresa materiais” e “contratação materiais”, apresentam menor oscilação nas de perturbação e acerto (τ_p, τ_a), enquanto os mais distantes, “serviços município” e “prestação município”, revelam fortes variações entre as frequências μ_e e ϕ_e, indicando forte dependência do modelo em relação a esses padrões.

No modelo SVM com WE (Figura 7), os padrões mais próximos do centroide no *cluster* 0 “contrato prestação” e “presente serviços município”, apresentam valores moderados de ρ e distribuição homogênea entre as classes, com variações discretas em κ_t e κ_e. Em contraste, os padrões mais distantes, “contratação município” e “serviços necessidade município”, possuem altos valores de ρ e discrepâncias significativas na razão de erro entre as classes,

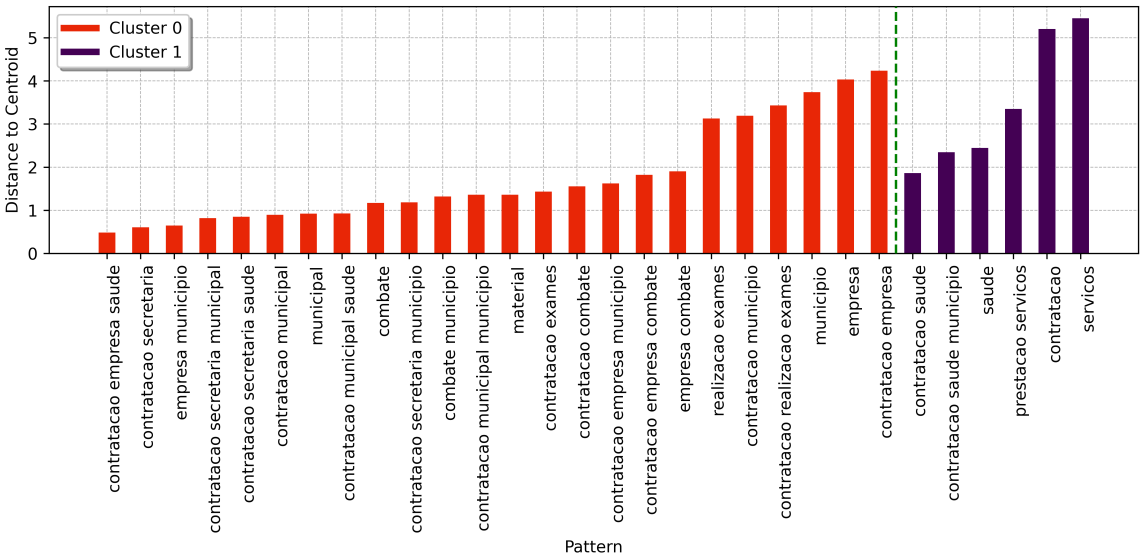
sugerindo que o modelo pode utilizá-los de forma enviesada. No *cluster* 1, os padrões mais próximos, “serviços” e “município”, exercem impacto moderado nas decisões do modelo, enquanto os mais distantes, “contratação” e “serviços município”, apresentam altos valores de ρ e variações expressivas em τ_p, indicando possível dependência excessiva do modelo em relação a esses padrões.

Figura 7 – SVM-WE-C0 – Gráfico da distância dos padrões aos centroides no agrupamento da classe 0 (base Contratos), SVM com *WE*. *Cluster* 0 em vermelho e *Cluster* 1 em roxo.



No modelo BERT (Figura 8), os padrões mais próximos do centroide no *cluster* 0, “contratação empresa saúde” e “contratação secretaria”, apresentam baixos valores de τ_p e τ_a, sugerindo que são menos determinantes para a classificação. Por outro lado, os padrões mais distantes, “contratação empresa” e “empresa”, possuem altos valores de ρ e grandes oscilações entre μ_t e ϕ_t, reforçando a hipótese de que se tratam de padrões espúrios. No *cluster* 1, os padrões mais próximos, “contratação saúde” e “contratação saúde município”, exibem distribuição equilibrada entre as classes. Em contraste, os mais distantes, “serviços” e “contratação”, apresentam as maiores valores em κ_t e κ_e, sugerindo que o modelo os utiliza de forma não causal.

Figura 8 – **BERT-C0** – Gráfico da distância dos padrões aos centroides no agrupamento da classe 0 (base Contratos), modelo BERTimbau. *Cluster* 0 em vermelho e *Cluster* 1 em roxo.



A análise das distâncias aos centroides revela que os padrões mais próximos apresentam distribuição equilibrada entre as classes e menor impacto nos erros do modelo. Em contraste, os padrões mais distantes exibem valores significativos nas métricas de frequência e maior influência nos erros e nas perturbações. A análise também mostra que os agrupamentos menores em cada modelo e representação textual concentram padrões com maior potencial de espuriedade, caracterizados por maior distância média ao centroide, maior peso global (ρ), variações mais expressivas entre κ_t, κ_e, e discrepância entre erro e perturbação. O padrão “contratação”, identificado nos menores agrupamentos, é altamente frequente no conjunto de dados. No entanto, sua distribuição assimétrica entre as classes leva o modelo a utilizá-lo como um atalho para a classificação. Além disso, seu alto peso global e grandes oscilações nas métricas de perturbação indicam que esse padrão se torna um fator determinante nas predições, mesmo quando não há uma relação semântica clara com a classe real da sentença.

5.2.2.3.1 Tendência Geral

Os padrões mais distantes do centroide demonstram desvios significativos em relação ao comportamento típico observado no agrupamento. Esses padrões apresentam maior variabilidade, com projeções extremas em métricas associadas às frequências e aos pesos globais. Consequentemente, possuem maior probabilidade de serem classificados como espúrios. Por outro lado, os padrões mais próximos do centroide mantêm maior consistência com as métricas médias do *cluster*, refletindo características típicas e menor variabilidade. A análise valida a hipótese, uma vez que as métricas de entrada indicam que padrões mais próximos apresentam menor grau de dependência, correspondendo a um risco reduzido de espuriedade em comparação com os padrões mais distantes.

5.2.3 Padrões Espúrios Potenciais Comuns aos Modelos

A hipótese central desta Tese sustenta que a distância entre os padrões e os centroides dos agrupamentos pode ser empregada como métrica para quantificar o grau de espuriedade. Padrões mais afastados tendem a se desviar do comportamento típico do agrupamento, indicando associações instáveis com a classe e, portanto, maior propensão a refletirem dependências espúrias aprendidas pelo modelo. Essa tendência foi confirmada empiricamente pelos resultados discutidos na seção anterior, nos quais visualizações baseadas em distância revelaram que os padrões associados a erros de classificação estavam frequentemente localizados nas extremidades dos agrupamentos. Com base nisso, analisamos os padrões identificados como potencialmente espúrios por, no mínimo, três combinações distintas de modelo e representação textual, por base de dados e classe.

A Tabela 10 apresenta os padrões mais recorrentes que se encontram distantes de seus centroides nos modelos aplicados ao conjunto de dados Contratos. Esses padrões aparecem em descrições contratuais variadas, o que os torna indicadores pouco confiáveis para distinção entre classes. Na classe 0, o termo “município” é frequentemente utilizado em contextos

Tabela 10 – Padrões potencialmente espúrios no conjunto de dados Contratos, identificados por ao menos três combinações de modelo e representação textual, agrupados por classe.

Classe	Padrão	Modelos que identificaram
Outras aquisições (0)	*município*[*], *serviços município*[†]	SVM (TF-IDF, WE), BERT
	contratação[‡], *contratação empresa*[§], *contratação município*[¶]	SVM (TF-IDF, WE), LRG (WE), BERT
	serviços[+]	SVM (TF-IDF, WE), LRG (TF-IDF, WE), BERT
Aquisições específicas de saúde (1)	*coronavirus covid*[–]	SVM (TF-IDF, WE), LRG (WE)
	fornecimento covid[@], *covid*[$], *aquisição covid*[&]	SVM (WE), LRG (WE), BERT
	aquisição[!]	SVM (TF-IDF, WE), LRG (TF-IDF, WE), BERT

administrativos desvinculados do conteúdo característico da classe, levando a erros de classificação. Da mesma forma, os termos “serviços”, “contratação” e “contratação empresa” ocorrem em diferentes tipos de contratos, confundindo os modelos. Na classe 1, “aquisição” é amplamente empregado em descrições de compras em diversas áreas, o que compromete sua estabilidade como atributo preditivo. Os termos “fornecimento” e “covid” aparecem de forma desproporcional em contratos da área da saúde, mas nem sempre definem o tipo contratual, gerando incertezas no processo de classificação.

No conjunto de dados Licitações, conforme mostrado na Tabela 11, os padrões “serviços” e “contratação” foram identificados como potencialmente espúrios na classe 0. Esses termos ocorrem de forma genérica em diferentes descrições de objetos licitatórios, levando a classificações incorretas. Por exemplo, a expressão “contratação empresa especializada” é utilizada tanto em processos de contratação de serviços quanto de aquisição de bens, tornando-se um termo pouco confiável para diferenciação entre classes. Na classe 1, os termos “aquisição” e “fornecimento referencia” aparecem frequentemente em múltiplos contextos, demonstrando sua instabilidade como atributos de classificação.

Tabela 11 – Padrões potencialmente espúrios no conjunto de dados *Licitações*, identificados por ao menos três combinações de modelo e representação textual, agrupados por classe.

Classe	Padrão	Modelos que identificaram
Contratação de serviços (0)	*contratação prestação*[*], *município*[#]	SVM (TF-IDF, WE), LRG (TF-IDF, WE)
	contratação especializada[†], *contratação empresa especializada*[‡]	SVM (WE), LRG (WE), BERT
	serviços[§], *contratação*[¶]	SVM (TF-IDF, WE), LRG (TF-IDF, WE), BERT
Aquisição de bens (1)	*fornecimento referencia*[+], *objeto fornecimento*[−]	SVM (TF-IDF, WE), LRG (WE), BERT
	aquisição[@], *fornecimento*[$]	SVM (TF-IDF, WE), LRG (TF-IDF, WE), BERT

Os resultados confirmam que padrões frequentes e genéricos apresentam maior potencial de serem espúrios, especialmente quando ocorrem em diferentes contextos e contribuem para erros do modelo. Assim como o termo "*spielberg*" em Wang e Culotta 2020b, os padrões "serviços", "aquisição" e "fornecimento" são padrões espúrios, pois, conforme o método proposto, aparecem tanto em sentenças semelhantes da mesma classe quanto da classe oposta, além de ocorrerem em sentenças incorretamente classificadas, o que mostra sua influencia em predições equivocadas. Esses termos surgem repetidamente em diferentes cenários, e o modelo os interpreta como fortes preditores, mesmo sem uma relação causal com a classe. Além disso, por estarem mais distantes dos centroides, conforme analisado em seções anteriores, esses padrões apresentam maior variabilidade, o que reforça essa tendência e aumenta o risco de associação espúria.

5.3 Experimento com o Conjunto de Dados IMDB e Comparação com Métodos de Referência

Essa Seção apresenta os resultados do método proposto utilizando o conjunto de dados IMDB, amplamente adotado como referência para classificação de sentimentos em textos. O conjunto contém 50.000 resenhas

de filmes extraídas da plataforma Internet Movie Database (IMDB), sendo 25.000 rotuladas como positivas e 25.000 como negativas (Maas et al. 2011). Este conjunto foi originalmente desenvolvido para treinar e testar modelos de análise de sentimento. Neste estudo, a abordagem proposta é aplicada para identificar padrões espúrios e medir sua robustez por meio da comparação com métodos de referência. Foram adotadas as mesmas configurações de modelos e representações textuais utilizadas nos experimentos com as bases de contratos e licitações, assegurando a consistência metodológica entre os cenários avaliados.

A Tabela 12 apresenta 25 padrões identificados como potencialmente espúrios para cada classe. A seleção inclui os cinco padrões mais distantes dos centroides dos agrupamentos e que foram detectados por pelo menos quatro modelos x representação textual, a coluna Distância representa a distância média dos padrões em relação ao centroide. De acordo com os resultados da etapa "Extrair Potenciais Padrões Espúrios", esses padrões influenciam os modelos a classificar sentenças de teste para a classe de importância, independentemente do contexto. Sua presença em sentenças de treinamento pode distorcer o processo de classificação, criando associações que não representam características semânticas genuínas do texto.

Tabela 12 – Padrões espúrios com maior distância em relação aos centroides dos agrupamentos para as classificações de sentimento negativo e positivo no conjunto de dados IMDB, detectados por $\geq$4 modelos x representação textual. Os padrões estão ordenados da esquerda para a direita e de cima para baixo dentro de cada agrupamento de classe, seguindo a ordem natural de leitura.

Classe Negativa						**Classe Positiva**					
Padrão	**Distância**	**Padrão**	**Distância**	**Padrão**	**Distância**	**Padrão**	**Distância**	**Padrão**	**Distância**	**Padrão**	**Distância**
bad	7.54	bad movie	6.52	worst	5.27	good	12.55	great	11.27	well	4.96
better	5.22	never	3.61	plot	3.49	great love	4.92	great well	4.80	excellent	4.60
nothing	3.37	boring	3.34	bad nothing	3.03	best good	4.14	best	3.97	love	3.87
bad plot	2.92	poor	2.90	bad movie plot	2.74	man	3.74	also	3.58	great time	3.44
director	2.61	movie plot	2.22	bad script	2.19	film story	3.32	good still	3.28	life	3.11
terrible	2.09	instead	2.07	movie never	1.54	also film	2.94	still	2.88	good see	2.84
problem	1.53	ridiculous	1.39	stupid	1.31	family	2.80	also story	2.79	good time	2.78
looks	1.24	money	1.19	acting plot	1.08	enjoyed	2.74	liked	2.70	true	2.65
movie	1.02					classic	2.61				

A análise desses padrões permite revelar vieses nos modelos de PLN e avaliar se determinadas palavras funcionam como atalhos de decisão, ignorando

o contexto completo da sentença. Para ilustrar, examinamos os termos "*movie*" e "*best*", identificados como espúrios por pelo menos quatro modelos.

5.3.0.1 Padrão Espúrio para a classe Negativa: "*Movie*"

O termo "movie" aparece com frequência nas resenhas como um elemento neutro em relação ao sentimento da resenha, referindo-se ao objeto da análise (o filme). No entanto, sua alta ocorrência em sentenças da classe 0 sugere que os modelos atribuem uma influência negativa indevida à sua presença.

Por exemplo, a sentença, classificada incorretamente na classe Negativa, "*I do not care if some people voted this movie to be bad. If you want the truth, this is a very good movie. It has everything a movie should have. You really should get this one."* expressa claramente uma opinião positiva sobre o filme, contradizendo a classificação negativa atribuída. Esse comportamento indica que o modelo correlaciona a mera presença da palavra "*movie*" com a classe Negativa, sem considerar a semântica global do texto. Adicionalmente, ao comparar "*movie*" com um padrão semelhante, "*movie plot*", observa-se que diversas sentenças foram rotuladas como Negativa apenas por conterem esse padrão, mesmo quando o contexto avaliava positivamente o enredo. Esse resultado sugere que os termos "*movie*" e "*plot*", quando combinados, não deveriam ser decisivos para a classificação.

Por exemplo, o texto: "*Note these comments are for people who have seen the movie. Vanilla Sky is a brilliant, complex, and thrilling movie that existentially explores exactly what the tagline says: love, hate, dreams, life, work, play, and friends. Maybe the movie plot can come into focus for confused moviegoers if one looks at it from a different angle, considering the following... This movie becomes a fascinating exploration of a human on a journey to find himself and what that means in today's pop culture society.*" foi rotulado Negativa apenas por conter "movie plot", apesar de conter elogios explícitos como "*brilliant*", *complex and thrilling movie*" e "*fascinating exploration*". Esse resultado confirma que a combinação "*movie plot*" pode atuar como um atalho para classificação negativa, mesmo quando não deveria desempenhar papel decisivo na decisão do modelo.

5.3.0.2 Padrão Espúrio para a classe Positiva: "*Best*"

O termo "*best*" geralmente aparece em resenhas positivas, sendo, portanto, esperado em sentenças da classe Positiva. No entanto, sua presençax pode levar a classificações incorretas quando usada fora de contexto. Por exemplo, a sentença "*The first part of Grease with John Travolta and Olivia Newton-John is one of the best movies for teens. This one is a very bad copy. The change is only in the sex. In the first one, the good one was Sandy. Here it is Michael. I prefer to watch the first Grease.*" contém o termo "*best*", mas expressa uma crítica negativa ao comparar desfavoravelmente o filme analisado com outro. A atribuição dessa sentença à classe Positiva sugere que o modelo considera apenas a presença do termo "*best*" e ignora o argumento completo.

A comparação com o padrão "*best good*" revela um comportamento semelhante. Muitas sentenças contendo essa combinação de palavras foram classificadas incorretamente como classe Positiva apenas por sua presença, mesmo quando o texto expressava uma crítica negativa. O trecho "*I debated quite a bit over what rating to give this one because it is my least favorite Herschell Gordon Lewis film so far other than The Gruesome Twosome, but it has the best acting I have seen in a Lewis film. However, we all know that is not saying much. Once the movie was done, I was happy because it felt like I had been sitting through a 4-hour movie, though it was only 82 min long. I am trying to see all of HGL's films, which is probably the only reason to see this one. The gore is good as usual—the one thing that Herschell seemed to get right. The acting is just as bad as usual with one exception. That exception is Frank Kress. Now, would I say that he is a good actor? No way. But he is good compared to everyone else. The story is boring and flat and goes nowhere, and by the end, I did not care what happened, just so long as it ended.*" contém as palavras "*best*" e "*good*", mas expressa uma avaliação predominantemente negativa do filme.

O autor destaca que, embora o filme apresente a "*best acting*" (melhor atuação) dentro da filmografia do diretor, isso não implica que a atuação seja realmente boa. Além disso, a crítica enfatiza que o filme é "*boring and flat*" (entediante e superficial), reforçando que a classificação positiva não reflete

a intenção real da resenha. Esse caso demonstra que a mera presença de termos associados à positividade pode levar o modelo a estabelecer correlações espúrias, resultando em classificações equivocadas.

5.4 Comparação com Métodos de Referência

Comparamos o método proposto com três abordagens de referência, previamente discutidas no Capítulo 2. A análise considerou os resultados dos trabalhos de Wang e Culotta 2020a, Wang e Culotta 2020b e Yadav et al. 2022, uma vez que esses estudos também utilizaram o conjunto de dados IMDB, o que possibilita uma comparação mais direta e consistente com os resultados obtidos por nossa abordagem. Wang e Culotta 2020a aplicam regressão logística combinada à *Closest Opposite Matching* para gerar contrafactuais via substituição de palavras por antônimos para analisar e mitigar padrões espúrios. Embora eficaz em aumentar a robustez do modelo, o método restringe-se à análise de termos isolados e depende de rotulação manual para validar causalidade e treinar o classificador com os contrafactuais gerados, em contraste com nossa proposta totalmente automatizada e capaz de identificar padrões compostos a partir de erros reais. Wang e Culotta 2020b tratam a espuriedade como uma tarefa de classificação binária de palavras, estimando seu impacto por meio do efeito de tratamento médio (ATE) em sentenças similares. Embora o método demonstre capacidade de generalização por meio da portabilidade do classificador entre diferentes domínios, requer um conjunto inicial rotulado manualmente e restringe-se a padrões unidimensionais, o que limita a complexidade dos padrões detectáveis e impõe maior esforço humano em comparação à nossa abordagem automatizada e multivariada. Por fim, Yadav et al. 2022 utilizam *Tsetlin Machine* (TM) para induzir regras lógicas compostas por literais afirmativos e negados, controladas por um parâmetro de especificidade. A abordagem oferece interpretabilidade e robustez a contrafactuais, mas está limitada à extração de padrões do tipo unigrama e exige dados manualmente modificados para treinar e validar as cláusulas, o que compromete sua escalabilidade frente à nossa proposta automatizada, multivariada e livre de intervenção manual.

A Tabela 13 detalha as diferenças metodológicas entre o método proposto e os métodos de referência, com base em três critérios centrais: (i) a abordagem adotada para identificação de padrões espúrios; (ii) a complexidade e natureza dos padrões considerados (unigramas ou combinações de palavras); e (iii) o grau de automação do processo, especialmente no que se refere à necessidade de intervenção manual. Esses critérios foram escolhidos por refletirem diretamente os principais objetivos e contribuições desta pesquisa, conforme discutido nos Capítulos 1 e 4.

Tabela 13 – Comparação entre o método proposto e abordagens de referência, aplicadas à base IMDB, considerando a abordagem adotada, o tipo de padrão encontrado e o grau de automação.

Critério	Método Proposto	Wang & Culotta (2020a)	Wang & Culotta (2020b)	Yadav et al. (2022)
Abordagem Proposta	XAI + Aprendizado não supervisionado	Substituição de palavras	Inferência causal por ATE	Regras lógicas via TM
Padrões	Até n palavras	Unigramas, generalização limitada	Unigramas, requer ajuste manual	Unigramas, sensível ao vocabulário
Esforço humano	Automatizado, sem intervenção manual	Requer validação manual	Necessita rotulagem inicial	Requer contrafactuais manuais

O grau de dependência de intervenção manual varia entre os métodos comparados. O método baseado em contrafactuais (Wang e Culotta 2020a) exige validação manual para garantir que as palavras substituídas por antônimos representem, de fato, padrões espúrios. A abordagem de inferência causal (Wang e Culotta 2020b) depende de anotadores humanos para rotular manualmente palavras influentes, o que introduz subjetividade ao processo. Já no método baseado em Máquinas de Tsetlin (Yadav et al. 2022), a construção manual de contrafactuais é essencial para refinar as regras lógicas aprendidas pelo modelo. O método proposto elimina essa necessidade ao automatizar a identificação e quantificação de padrões espúrios. A explicabilidade baseada em XAI permite extrair palavras importantes diretamente dos modelos, enquanto a análise de perturbações e a clusterização quantificam a influência dos padrões. Essa abordagem torna a detecção de padrões espúrios mais escalável e adaptável a diferentes domínios. No entanto, como em qualquer estudo científico,

a análise dos resultados pode envolver especialistas do domínio para interpretar os padrões detectados e validar os achados mais relevantes. Essa etapa, contudo, não constitui uma dependência operacional do método, mas sim uma prática recomendada para garantir a robustez e a aplicabilidade das conclusões em contextos específicos.

Além das palavras comuns identificadas pelo nosso método (ver Tabela 14), apenas a abordagem proposta foi capaz de detectar padrões compostos, tais como "*movie plot*" (enredo filme) e "*bad movie plot*" (enredo ruim filme). A identificação desses padrões compostos representa uma vantagem substancial na análise de dependências que os modelo pode aprender baseadas em coocorrência, em vez de relações semânticas mais amplas.

Tabela 14 – Padrões identificados tanto pelo método proposto quanto por métodos de referência. "Classificação" refere-se à atribuição feita pelos respectivos trabalhos.

Referência	**Padrões**	**Classificação**
Wang e Culotta (2020b)	*boring*	Espúrio
Wang e Culotta (2020b)	*stupid, best, enjoyed*	Genuíno
Wang e Culotta (2020a)	*movie, best*	Espúrio
Wang e Culotta (2020a)	*terrible, excellent, great*	Genuíno
Yadav et al. (2022)	*best, instead, money, plot, worst, family*	Espúrio

Por exemplo, a presença do padrão "bad movie plot" pode indicar que o modelo está superdependente de composições recorrentes entre termos altamente polarizados ("bad") e expressões neutras ("movie plot") para predizer a classe negativa. Embora semanticamente plausível, esse tipo de atalho ignora o conteúdo semântico completo da sentença e favorece decisões baseadas em padrões locais recorrentes, o que compromete a generalização. A identificação desses padrões compostos por nosso método revela esse viés estrutural, ao capturar padrões que se afastam dos centroides e recorrentes em erros de classificação.

5.5 Considerações Finais

Os experimentos demonstraram a eficácia do método proposto na detecção e quantificação automática de padrões espúrios em tarefas de classificação binária. A aplicação aos conjuntos Contratos, Licitações e IMDB revelou que diferentes modelos e representações textuais — como SVM, Regressão Logística e BERTimbau — são suscetíveis a termos recorrentes e semanticamente genéricos que influenciam indevidamente suas decisões. Por meio da combinação entre explicabilidade, recuperação de sentenças semelhantes, perturbações controladas e agrupamento não supervisionado, foi possível evidenciar como certos padrões funcionam como atalhos, mesmo em arquiteturas robustas como o BERT.

A abordagem mostrou-se generalizável a diferentes domínios e vantajosa frente a métodos de referência, ao dispensar contrafactuais ou listas prévias de potenciais padrões espúrios. A métrica baseada na distância ao centroide de agrupamentos lógicos destacou-se como estimador contínuo de espuriedade, evitando dicotomias artificiais e permitindo avaliar nuances contextuais — uma solução mais compatível com a natureza probabilística dos modelos modernos.

O desempenho superior do BERT ao capturar padrões mais sutis deve ser interpretado com cautela. Embora o modelo ofereça ganhos em cobertura semântica, ele também carrega vieses dos dados de treinamento e apresenta custo computacional elevado. Além disso, sua estrutura de tokenização impede, na configuração atual, a contabilização de padrões semanticamente equivalentes com vocabulário divergente. O método não apenas identifica padrões espúrios, mas estima sua confiabilidade e impacto sobre as decisões do modelo, contribuindo de forma inovadora para o diagnóstico automatizado de viés em PLN.

Em síntese, os experimentos demonstraram a eficácia do método proposto na detecção e quantificação de padrões espúrios, tanto em dados de domínio específico (Contratos e Licitações) quanto em dados de referência amplamente utilizados (IMDB). A análise comparativa revelou que diferentes modelos e representações textuais são suscetíveis a atalhos espúrios, mas que

o método proposto oferece métricas contínuas e interpretações visuais para caracterizar tais padrões. Esses resultados consolidam a validade da abordagem e abrem espaço para uma reflexão crítica sobre suas contribuições, limitações e possíveis extensões. O próximo capítulo discute essas implicações, sintetizando conclusões, destacando restrições metodológicas e apontando direções para trabalhos futuros.

6 Conclusões, Discussão, Limitações e Trabalhos Futuros

6.1 Introdução

Este capítulo apresenta a síntese dos resultados obtidos, explicita as principais limitações do estudo e propõe direções para trabalhos futuro. A Tese propôs um método agnóstico ao modelo, baseado na integração entre técnicas XAI e aprendizado não supervisionado, voltado à detecção, interpretação e quantificação de padrões espúrios em tarefas de classificação binária em PLN. A abordagem foi aplicada a dados reais do TCE-PI e posteriormente validada na base IMDB, com comparação a métodos de referência. As contribuições teóricas, metodológicas e empíricas deste trabalho foram apresentadas na Seção 1.3 do Capítulo 1, de modo a contextualizar desde o início os elementos inovadores da proposta.

6.2 Conclusões

A análise dos resultados revela que os modelos avaliados exibem diferentes graus de sensibilidade a padrões compostos. O modelo BERT obteve melhor desempenho na detecção de padrões espúrios, graças à sua capacidade de captar dependências semânticas de maior ordem. No entanto, o método proposto permitiu que modelos lineares, com representações menos ricas, fossem eficientes na identificação de padrões complexos e nuances contextuais.

A métrica de distância ao centroide demonstrou utilidade para quantificar o grau de espuriedade dos padrões. Padrões reconhecidos pelos especialistas como irrelevantes ou genéricos, como “serviços”, “aquisição”, “empresa” e “objeto edital”, frequentemente aparecem como *outliers* nos *clusters*, confirmando a hipótese central da pesquisa.

Além disso, observou-se que os *clusters* gerados pelos diferentes modelos refletem as peculiaridades de cada representação textual. Apesar de os agrupamentos formados por BERT apresentarem maior consistência interna, os demais modelos também forneceram agrupamentos coerentes, quando avaliados à luz dos padrões identificados.

A intervenção manual foi limitada à definição de parâmetros técnicos do método (como o número de *clusters* e limiares de corte), sem necessidade de rotulação manual ou criação de instâncias artificiais. Esse aspecto diferencia o método de abordagens tradicionais e reforça sua capacidade de escalabilidade.

Dessa forma, os objetivos delineados nesta pesquisa foram atingidos: desenvolvemos um método automatizado e agnóstico ao modelo para identificar padrões espúrios; validamos a métrica baseada na distância ao centroide como estimador da espuriedade; e demonstramos a aplicabilidade prática da abordagem em bases reais e de referência. Os resultados reforçam o potencial do método como ferramenta de apoio ao diagnóstico e mitigação de padrões espúrios.

6.3 Limitações

Apesar dos resultados promissores, o método proposto apresenta limitações que devem ser consideradas. Sua confiabilidade depende da estabilidade dos agrupamentos gerados e da precisão das explicações produzidas, dois fatores que impõem desafios conceituais e práticos. Algumas decisões técnicas adotadas, como o uso do algoritmo *K-means* e do explicador LIME, introduzem restrições que podem afetar a robustez e a aplicabilidade da abordagem em diferentes contextos. Além disso, há limitações relacionadas à representação semântica dos padrões, ao custo computacional envolvido e à necessidade de validação por especialistas. A seguir, descrevem-se as principais limitações observadas:

- **Dependência geométrica do *K-means*:** A estrutura esférica presumida pelo *K-means* pode distorcer agrupamentos reais, especialmente em representações de alta dimensionalidade com densidade variável.

- **Fragilidade dos explicadores:** Técnicas como LIME e coeficientes lineares são sensíveis a ruídos e instabilidade local. Explicadores alternativos, como SHAP, podem ser explorados, mas impõem desafios de custo computacional e interpretação.
- **Limitação na contagem de padrões com BERT:** O método considera apenas padrões textuais idênticos, ignorando equivalências semânticas identificadas pelo BERT. Assim, expressões como "aquisição de materiais hospitalares" e "compra de suprimentos hospitalares" são tratadas como distintas, o que pode fragmentar padrões semanticamente redundantes e limitar o aproveitamento do modelo.
- **Custo computacional:** Os testes utilizando o BERT com LIME demandaram mais recursos computacionais, o que pode limitar sua aplicação em cenários operacionais mais restritos. Avaliamos o Explicador *Transformers Interpret* (Lal et al. 2021), porém, sua tokenização por subpalavras (e.g., "aquisição" $\rightarrow$ "aqui", "##sição") não atende aos critérios da Equação 4.6.
- **Validação ainda dependente de especialistas:** A interpretação final dos padrões e agrupamentos depende da expertise de avaliadores humanos, o que pode limitar a escalabilidade da abordagem em contextos sem disponibilidade de domínio.

6.4 Trabalhos Futuros

Os resultados obtidos abrem espaço para aprimoramentos e extensões da abordagem proposta. Algumas limitações identificadas ao longo do estudo apontam diretamente para direções de pesquisa que podem ampliar a robustez, a eficiência e a aplicabilidade do método em contextos mais diversos. Além disso, novas possibilidades emergem a partir da análise crítica dos experimentos, sugerindo melhorias tanto na modelagem quanto na interação com especialistas. A seguir, são elencadas propostas concretas para investigações futuras:

1. **Substituição do *K-means*:** Avaliar alternativas mais robustas como DBSCAN, HDBSCAN, GMM ou agrupamento hierárquico, capazes de identificar *clusters* com formatos arbitrários e densidades variáveis.

2. **Unificação semântica de padrões:** Incorporar técnicas baseadas em *embeddings* contextuais, ontologias ou substituições controladas para agrupar expressões semanticamente equivalentes, ampliando a sensibilidade do método.

3. **Otimização computacional do método:** Investigar alternativas que reduzam a complexidade computacional do método, especialmente no uso de explicadores. Nos modelos lineares, o SHAP (SHapley Additive exPlanations) surge como alternativa promissora ao LIME, pois atende aos requisitos da Equação 4.6 e pode oferecer explicações igualmente robustas com menor custo. Para o BERT, futuras abordagens devem buscar explicadores compatíveis que superem as limitações do *Transformers Interpret*.

4. **Interfaces interativas para especialistas:** Desenvolver sistemas visuais que reduzam o esforço na validação de padrões por especialistas, permitindo análise dinâmica e baseada em visualizações intuitivas.

5. **Extensão para padrões mais longos:** Ampliar a janela de análise para padrões com mais de três palavras, explorando relações sintáticas e semânticas de maior complexidade nos textos.

6. **Transferência para dados estruturados:** Avaliar o uso da abordagem em bases tabulares, especialmente em tarefas de seleção de atributos e auditoria explicável de modelos, expandindo o escopo da aplicabilidade.

Referências

ADADI, A.; BERRADA, M. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). **IEEE access**, IEEE, v. 6, p. 52138–52160, 2018. Citado na página 28.

ALI, S.; ABUHMED, T.; EL-SAPPAGH, S.; MUHAMMAD, K.; ALONSO-MORAL, J. M.; CONFALONIERI, R.; GUIDOTTI, R.; SER, J. D.; DÍAZ-RODRÍGUEZ, N.; HERRERA, F. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. **Information fusion**, Elsevier, v. 99, p. 101805, 2023. Citado 5 vezes nas páginas 20, 27, 28, 57 e 69.

ANDERS, C. J.; WEBER, L.; NEUMANN, D.; SAMEK, W.; MÜLLER, K.-R.; LAPUSCHKIN, S. Finding and removing clever hans: Using explanation methods to debug and improve deep models. **Information Fusion**, Elsevier, v. 77, p. 261–295, 2022. Citado na página 20.

ARRIETA, A. B.; DÍAZ-RODRÍGUEZ, N.; SER, J. D.; BENNETOT, A.; TABIK, S.; BARBADO, A.; GARCÍA, S.; GIL-LÓPEZ, S.; MOLINA, D.; BENJAMINS, R. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. **Information fusion**, Elsevier, v. 58, p. 82–115, 2020. Citado na página 19.

BARLOW, H. B. Unsupervised learning. **Neural computation**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 1, n. 3, p. 295–311, 1989. Citado na página 29.

BHOLOWALIA, P.; KUMAR, A. Ebk-means: A clustering technique based on elbow method and k-means in wsn. **International Journal of Computer Applications**, Citeseer, v. 105, n. 9, 2014. Citado na página 30.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. Enriching word vectors with subword information. **Transactions of the association for computational linguistics**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 5, p. 135–146, 2017. Citado na página 27.

BOUNTAKAS, P.; KOUTROUMPOUCHOS, K.; XENAKIS, C. A comparison of natural language processing and machine learning methods for phishing email detection. In: **Proceedings of the 16th International Conference on Availability, Reliability and Security**. [S.l.: s.n.], 2021. p. 1–12. Citado na página 24.

BREIMAN, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). **Statistical science**, Institute of Mathematical Statistics, v. 16, n. 3, p. 199–231, 2001. Citado na página 28.

BRO, R.; SMILDE, A. K. Principal component analysis. **Analytical methods**, Royal Society of Chemistry, v. 6, n. 9, p. 2812–2831, 2014. Citado na página 30.

CAMPELLO, R. J.; MOULAVI, D.; SANDER, J. Density-based clustering based on hierarchical density estimates. In: SPRINGER. **Pacific-Asia conference on knowledge discovery and data mining**. [S.l.], 2013. p. 160–172. Citado na página 66.

CARDOZO, S.; MONTERO, G. I.; KAZHDAN, D.; DIMANOV, B.; WIJAYA, M.; JAMNIK, M.; LIO, P. Explainer divergence scores (eds): Some post-hoc explanations may be effective for detecting unknown spurious correlations. **arXiv preprint arXiv:2211.07650**, 2022. Citado na página 20.

CHANG, H.; HE, Z.; LIU, C.; TAO, S.; XU, J.; LI, A. C.; ZENG, Y. Statistical learning based treatment effect calculation for spurious classifiers. In: IEEE. **2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)**. [S.l.], 2022. p. 457–463. Citado 2 vezes nas páginas 39 e 48.

CHEN, B.; CARVALHO, W.; BARACALDO, N.; LUDWIG, H.; EDWARDS, B.; LEE, T.; MOLLOY, I.; SRIVASTAVA, B. Detecting backdoor attacks on deep neural networks by activation clustering. **arXiv preprint arXiv:1811.03728**, 2018. Citado na página 17.

CHEW, O.; LIN, H.-T.; CHANG, K.-W.; HUANG, K.-H. Understanding and mitigating spurious correlations in text classification with neighborhood analysis. In: GRAHAM, Y.; PURVER, M. (Ed.). **Findings of the Association for Computational Linguistics: EACL 2024**. St. Julian's, Malta: Association for Computational Linguistics, 2024. p. 1013–1025. Disponível em: <https://aclanthology.org/2024.findings-eacl.68>. Citado 3 vezes nas páginas 20, 44 e 48.

CHOPRA, A.; PRASHAR, A.; SAIN, C. Natural language processing. **International journal of technology enhancements and emerging engineering research**, Citeseer, v. 1, n. 4, p. 131–134, 2013. Citado na página 23.

CHOU, Y.-L.; MOREIRA, C.; BRUZA, P.; OUYANG, C.; JORGE, J. Counterfactuals and causability in explainable artificial intelligence: Theory,

algorithms, and applications. **Information Fusion**, Elsevier, v. 81, p. 59–83, 2022. Citado 2 vezes nas páginas 20 e 33.

COHN, D.; ATLAS, L.; LADNER, R. Improving generalization with active learning. **Machine learning**, Springer, v. 15, p. 201–221, 1994. Citado na página 51.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018. Citado 3 vezes nas páginas 25, 27 e 67.

DU, Y.; YAN, J.; CHEN, Y.; LIU, J.; ZHAO, S.; SHE, Q.; WU, H.; WANG, H.; QIN, B. Less learn shortcut: Analyzing and mitigating learning of spurious feature-label correlation. **arXiv preprint arXiv:2205.12593**, 2022. Citado na página 20.

EISENSTEIN, J. Natural language processing. **Jacob Eisenstein**, 2018. Citado na página 23.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **kdd**. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231. Citado na página 66.

FONSECA, E.; ROSA, J. L. G. Mac-morpho revisited: Towards robust part-of-speech tagging. In: **Proceedings of the 9th Brazilian symposium in information and human language technology**. [S.l.: s.n.], 2013. Citado na página 51.

GAN, G. A k-means algorithm with automatic outlier detection. **Electronics**, MDPI, v. 14, n. 9, p. 1723, 2025. Citado 3 vezes nas páginas 65, 70 e 79.

GAN, G.; NG, M. K.-P. K-means clustering with outlier removal. **Pattern Recognition Letters**, Elsevier, v. 90, p. 8–14, 2017. Citado 3 vezes nas páginas 65, 70 e 79.

GAURAV, D.; TIWARI, S. Interpretability vs explainability: The black box of machine learning. In: IEEE. **2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)**. [S.l.], 2023. p. 523–528. Citado na página 28.

GAUTAM, S.; HÖHNE, M. M.-C.; HANSEN, S.; JENSSEN, R.; KAMPFFMEYER, M. This looks more like that: Enhancing self-explaining models by prototypical relevance propagation. **Pattern Recognition**, Elsevier, v. 136, p. 109172, 2023. Citado 2 vezes nas páginas 17 e 20.

GHAHRAMANI, Z. Unsupervised learning. In: **Summer school on machine learning**. [S.l.]: Springer, 2003. p. 72–112. Citado na página 29.

GHOSAL, S. S.; LI, Y. Distributionally robust optimization with probabilistic group. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. [S.l.: s.n.], 2023. v. 37, n. 10, p. 11809–11817. Citado 2 vezes nas páginas 46 e 48.

GOLDSTEIN, A.; KAPELNER, A.; BLEICH, J.; PITKIN, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. **journal of Computational and Graphical Statistics**, Taylor & Francis, v. 24, n. 1, p. 44–65, 2015. Citado na página 28.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: MIT press, 2016. Citado na página 24.

HARRELL, F. E. et al. **Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis**. [S.l.]: Springer, 2001. v. 608. Citado na página 24.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H.; FRIEDMAN, J. H. **The elements of statistical learning: data mining, inference, and prediction**. [S.l.]: Springer, 2009. v. 2. Citado 2 vezes nas páginas 29 e 52.

HE, Z.; XU, X.; DENG, S. Discovering cluster-based local outliers. **Pattern recognition letters**, Elsevier, v. 24, n. 9-10, p. 1641–1650, 2003. Citado na página 70.

HOWELL, K.; BARNES, M.; CURTIS, J. R.; ENGELBERG, R. A.; LEE, R. Y.; LOBER, W. B.; SIBLEY, J.; COHEN, T. Controlling for confounding variables: accounting for dataset bias in classifying patient-provider interactions. **Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability**, Springer, p. 271–282, 2021. Citado na página 18.

HU, Z. F.; KUFLIK, T.; MOCANU, I. G.; NAJAFIAN, S.; TAL, A. S. Recent studies of xai-review. In: **Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization**. [S.l.: s.n.], 2021. p. 421–431. Citado na página 28.

HUANG, C.-H.; YIN, J.; HOU, F. A text similarity measurement combining word semantic information with tf-idf method. **Jisuanji Xuebao(Chinese Journal of Computers)**, Science Press(Beijing),| a 16 Donghuangchenggen North Street| c Beijing| z . . . , v. 34, n. 5, p. 856–864, 2011. Citado na página 68.

HUANG, T.; WANG, S.; ZHU, W. An adaptive kernelized rank-order distance for clustering non-spherical data with high noise. **International Journal of Machine**

Learning and Cybernetics, Springer, v. 11, n. 8, p. 1735–1747, 2020. Citado 3 vezes nas páginas 65, 70 e 79.

IZMAILOV, P.; KIRICHENKO, P.; GRUVER, N.; WILSON, A. G. On feature learning in the presence of spurious correlations. **Advances in Neural Information Processing Systems**, v. 35, p. 38516–38532, 2022. Citado na página 20.

IZZA, Y.; HUANG, X.; IGNATIEV, A.; NARODYTSKA, N.; COOPER, M.; MARQUES-SILVA, J. On computing probabilistic abductive explanations. **International Journal of Approximate Reasoning**, Elsevier, v. 159, p. 108939, 2023. Citado na página 28.

JAKOBSEN, T. S. T.; BARRETT, M.; SøGAARD, A. Spurious correlations in cross-topic argument mining. In: **Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics**. [S.l.]: Association for Computational Linguistics, 2021. p. 263–277. Tenth Joint Conference on Lexical and Computational Semantics - SEM 2021 ; Conference date: 05-08-2021 Through 06-08-2021. Citado 2 vezes nas páginas 43 e 48.

JOSHI, N.; PAN, X.; HE, H. Are all spurious features in natural language alike? an analysis through a causal lens. **arXiv preprint arXiv:2210.14011**, 2022. Citado 2 vezes nas páginas 39 e 48.

KAMBHATLA, N.; LEEN, T. K. Dimension reduction by local principal component analysis. **Neural computation**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 9, n. 7, p. 1493–1516, 1997. Citado na página 30.

KATTAKINDA, P.; FEIZI, S. Focus: Familiar objects in common and uncommon settings. **arXiv preprint arXiv:2110.03804**, 2021. Citado na página 20.

KEITH, K. A.; JENSEN, D.; O'CONNOR, B. Text and causal inference: A review of using text to remove confounding from causal estimates. **arXiv preprint arXiv:2005.00649**, 2020. Citado na página 18.

KLEINBAUM, D. G.; DIETZ, K.; GAIL, M.; KLEIN, M.; KLEIN, M. **Logistic regression**. [S.l.]: Springer, 2002. Citado na página 24.

KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: MONTREAL, CANADA. **Ijcai**. [S.l.], 1995. v. 14, n. 2, p. 1137–1145. Citado 2 vezes nas páginas 49 e 52.

KUMAR, A.; DESHPANDE, A.; SHARMA, A. Causal effect regularization: Automated detection and removal of spurious correlations. In: OH, A.; NEUMANN, T.; GLOBERSON, A.; SAENKO, K.; HARDT, M.; LEVINE, S. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2023. v. 36, p. 20942–20984. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2023/file/42770daf4a3384b712ea9c36e9279998-Paper-Conference.pdf>. Citado na página 20.

LAL, V.; MA, A.; AFLALO, E.; HOWARD, P.; SIMOES, A.; KORAT, D.; PEREG, O.; SINGER, G.; WASSERBLAT, M. Interpret: An interactive visualization tool for interpreting transformers. In: **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations**. [S.l.: s.n.], 2021. p. 135–142. Citado na página 96.

LAMPRIDIS, O.; STATE, L.; GUIDOTTI, R.; RUGGIERI, S. Explaining short text classification with diverse synthetic exemplars and counter-exemplars. **Machine learning**, Springer, v. 112, n. 11, p. 4289–4322, 2023. Citado 2 vezes nas páginas 44 e 48.

LAPUSCHKIN, S.; WÄLDCHEN, S.; BINDER, A.; MONTAVON, G.; SAMEK, W.; MÜLLER, K.-R. Unmasking clever hans predictors and assessing what machines really learn. **Nature communications**, Nature Publishing Group UK London, v. 10, n. 1, p. 1096, 2019. Citado 2 vezes nas páginas 17 e 18.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: PMLR. **International conference on machine learning**. [S.l.], 2014. p. 1188–1196. Citado na página 27.

LI, X.; ORABONA, F. On the convergence of stochastic gradient descent with adaptive stepsizes. In: PMLR. **The 22nd international conference on artificial intelligence and statistics**. [S.l.], 2019. p. 983–992. Citado na página 24.

LIMA, W.; SILVA, V.; SILVA, J.; LIRA, R.; PAIVA, A. Bidcorpus: A multifaceted learning dataset for public procurement. **Data in Brief**, Elsevier, v. 58, p. 111202, 2025. Citado na página 51.

LIU, C.; GAN, L.; KUANG, K.; WU, F. Investigating the robustness of natural language generation from logical forms via counterfactual samples. **arXiv preprint arXiv:2210.08548**, 2022. Citado 2 vezes nas páginas 36 e 48.

MAAS, A.; DALY, R. E.; PHAM, P. T.; HUANG, D.; NG, A. Y.; POTTS, C. Learning word vectors for sentiment analysis. In: **Proceedings of the 49th annual**

meeting of the association for computational linguistics: Human language technologies. [S.l.: s.n.], 2011. p. 142–150. Citado na página 86.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. [S.l.], 1967. v. 1, n. 14, p. 281–297. Citado 2 vezes nas páginas 29 e 30.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013. Citado na página 27.

MING, Y.; YIN, H.; LI, Y. On the impact of spurious correlation for out-of-distribution detection. In: **Proceedings of the AAAI conference on artificial intelligence**. [S.l.: s.n.], 2022. v. 36, n. 9, p. 10051–10059. Citado 2 vezes nas páginas 41 e 48.

MU, S.; LI, Y.; ZHAO, W. X.; WANG, J.; DING, B.; WEN, J.-R. Alleviating spurious correlations in knowledge-aware recommendations through counterfactual generator. In: **Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval**. [S.l.: s.n.], 2022. p. 1401–1411. Citado na página 18.

MUSSABAYEV, R.; MLADENOVIC, N.; JARBOUI, B.; MUSSABAYEV, R. How to use k-means for big data clustering? **Pattern Recognition**, Elsevier, v. 137, p. 109269, 2023. Citado na página 29.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1532–1543. Citado na página 27.

PEZESHKPOUR, P.; JAIN, S.; SINGH, S.; WALLACE, B. C. Combining feature and instance attribution to detect artifacts. 7 2021. Disponível em: <http://arxiv.org/abs/2107.00323>. Citado 2 vezes nas páginas 18 e 49.

PLUMB, G.; RIBEIRO, M. T.; TALWALKAR, A. Finding and fixing spurious patterns with explanations. **arXiv preprint arXiv:2106.02112**, 2021. Citado 3 vezes nas páginas 20, 41 e 48.

RADFORD, A.; WU, J.; CHILD, R.; LUAN, D.; AMODEI, D.; SUTSKEVER, I. et al. Language models are unsupervised multitask learners. **OpenAI blog**, v. 1, n. 8, p. 9, 2019. Citado na página 27.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. **arXiv preprint arXiv:1908.10084**, 2019. Citado na página 67.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should I trust you?": Explaining the predictions of any classifier. **CoRR**, abs/1602.04938, 2016. Disponível em: <http://arxiv.org/abs/1602.04938>. Citado 2 vezes nas páginas 28 e 69.

SCHWARTZ, R.; STANOVSKY, G. On the limitations of dataset balancing: The lost battle against spurious correlations. 4 2022. Disponível em: <http://arxiv.org/abs/2204.12708>. Citado 2 vezes nas páginas 18 e 20.

SENGUPTA, P.; ZHANG, Y.; MAHARJAN, S.; ELIASSEN, F. Balancing explainability-accuracy of complex models. **arXiv preprint arXiv:2305.14098**, 2023. Citado na página 28.

SERRANO, S.; DODGE, J.; SMITH, N. A. Stubborn lexical bias in data and models. **arXiv preprint arXiv:2306.02190**, 2023. Citado 2 vezes nas páginas 40 e 48.

SHAO, Y.; TAYLOR, S.; MARSHALL, N.; MORIOKA, C.; ZENG-TREITLER, Q. Clinical text classification with word embedding features vs. bag-of-words features. In: IEEE. **2018 IEEE International conference on big data (big data)**. [S.l.], 2018. p. 2874–2878. Citado na página 26.

SHIRNIN, A.; ANDREEV, N.; POTAPOVA, S.; ARTEMOVA, E. Analyzing the robustness of vision & language models. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, IEEE, 2024. Citado 2 vezes nas páginas 42 e 48.

SOARES, H.; VERAS, R.; MOURA, R.; PAIVA, A. Using explainability to find spurious patterns in textual datasets. In: ABRAHAM, A.; BAJAJ, A.; HANNE, T.; SIARRY, P. (Ed.). **Intelligent Systems Design and Applications**. Cham: Springer Nature Switzerland, 2024. p. 424–434. ISBN 978-3-031-64779-6. Citado 3 vezes nas páginas 25, 63 e 69.

SOUSA, R. C. C. de; LOPES, H. Portuguese pos tagging using blstm without handcrafted features. In: SPRINGER. **Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 24th Iberoamerican Congress, CIARP 2019, Havana, Cuba, October 28-31, 2019, Proceedings 24**. [S.l.], 2019. p. 120–130. Citado na página 51.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: SPRINGER. **Brazilian conference on intelligent systems**. [S.l.], 2020. p. 403–417. Citado na página 67.

SRIVASTAVA, M. Mitigating spurious correlations in machine learning models: Techniques and applications. **OSF Preprints. April**, v. 21, 2023. Citado na página 20.

SYAKUR, M.; KHOTIMAH, B. K.; ROCHMAN, E.; SATOTO, B. D. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In: IOP PUBLISHING. **IOP conference series: materials science and engineering**. [S.l.], 2018. v. 336, p. 012017. Citado na página 30.

TJANDRA, D.; WIENS, J. Leveraging an alignment set in tackling instance-dependent label noise. In: PMLR. **Conference on Health, Inference, and Learning**. [S.l.], 2023. p. 477–497. Citado na página 20.

VALE, A. H.; SANTOS, P.; SOARES, H.; MOURA, R. S. Automatic classification of public expenses in the fight against covid-19: A case study of tce/pi. In: **Proceedings of the XIX Brazilian Symposium on Information Systems**. New York, NY, USA: Association for Computing Machinery, 2023. (SBSI '23), p. 221–228. ISBN 9798400707599. Disponível em: <https://doi.org/10.1145/3592813.3592909>. Citado na página 50.

VEITCH, V.; D'AMOUR, A.; YADLOWSKY, S.; EISENSTEIN, J. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. **arXiv preprint arXiv:2106.00545**, 2021. Citado 2 vezes nas páginas 36 e 48.

WANG, H.; WANG, Z.; DU, M.; YANG, F.; ZHANG, Z.; DING, S.; MARDZIEL, P.; HU, X. Score-cam: Score-weighted visual explanations for convolutional neural networks. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops**. [S.l.: s.n.], 2020. p. 24–25. Citado na página 28.

WANG, T.; SRIDHAR, R.; YANG, D.; WANG, X. Identifying and mitigating spurious correlations for improving robustness in nlp models. **arXiv preprint arXiv:2110.07736**, 2021. Citado 4 vezes nas páginas 19, 37, 43 e 48.

WANG, Z.; CULOTTA, A. Robustness to spurious correlations in text classification via automatically generated counterfactuals. 12 2020a. Disponível em: <http://arxiv.org/abs/2012.10040>. Citado 5 vezes nas páginas 33, 43, 48, 89 e 90.

WANG, Z.; CULOTTA, A. Identifying spurious correlations for robust text classification. 10 2020b. Disponível em: <http://arxiv.org/abs/2010.02458>. Citado 9 vezes nas páginas 18, 20, 36, 43, 48, 49, 85, 89 e 90.

WIJNHOVEN, R. G.; WITH, P. de. Fast training of object detection using stochastic gradient descent. In: IEEE. **2010 20th International conference on pattern recognition**. [S.l.], 2010. p. 424–427. Citado na página 25.

WU, J.; HOOI, B. Probing spurious correlations in popular event-based rumor detection benchmarks. In: SPRINGER. **Joint European Conference on Machine Learning and Knowledge Discovery in Databases**. [S.l.], 2022. p. 274–290. Citado 2 vezes nas páginas 42 e 48.

WU, T.; DING, X.; TANG, M.; ZHANG, H.; QIN, B.; LIU, T. Noisywikihow: A benchmark for learning with real-world noisy labels in natural language processing. **arXiv preprint arXiv:2305.10709**, 2023. Citado 2 vezes nas páginas 20 e 54.

WU, Y.; GARDNER, M.; STENETORP, P.; DASIGI, P. Generating data to mitigate spurious correlations in natural language inference datasets. **arXiv preprint arXiv:2203.12942**, 2022. Citado 2 vezes nas páginas 34 e 48.

YADAV, R. K.; LEI, J.; GRANMO, O.-C.; GOODWIN, M. Robust interpretable text classification against spurious correlations using and-rules with negation. In: INTERNATIONAL JOINT CONFERENCES ON ARTIFICIAL INTELLIGENCE. **IJCAI International Joint Conference on Artificial Intelligence**. [S.l.], 2022. Citado 4 vezes nas páginas 35, 48, 89 e 90.

YANG, M.; ZHANG, X.; WANG, J.; ZHOU, X. Causal representation for few-shot text classification. **Applied Intelligence**, Springer, v. 53, n. 18, p. 21422–21432, 2023. Citado 2 vezes nas páginas 45 e 48.

YUAN, L.; CHEN, Y.; CUI, G.; GAO, H.; ZOU, F.; CHENG, X.; JI, H.; LIU, Z.; SUN, M. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. In: OH, A.; NEUMANN, T.; GLOBERSON, A.; SAENKO, K.; HARDT, M.; LEVINE, S. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2023. v. 36, p. 58478–58507. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2023/file/b6b5f50a2001ad1cbccca96e693c4ab4-Paper-Datasets_and_Benchmarks.pdf>. Citado na página 17.

ZAFAR, M. R.; KHAN, N. Deterministic local interpretable model-agnostic explanations for stable explainability. **Machine Learning and Knowledge Extraction**, MDPI, v. 3, n. 3, p. 525–541, 2021. Citado na página 28.

ZHANG, T.; KISHORE, V.; WU, F.; WEINBERGER, K. Q.; ARTZI, Y. Bertscore: Evaluating text generation with bert. **arXiv preprint arXiv:1904.09675**, 2019. Citado na página 68.

ZHANG, Y.; PAN, L.; TAN, S.; KAN, M.-Y. Interpreting the robustness of neural nlp models to textual perturbations. **arXiv preprint arXiv:2110.07159**, 2021. Citado 2 vezes nas páginas 42 e 48.

ZHOU, Y.; XU, P.; LIU, X.; AN, B.; AI, W.; HUANG, F. Explore spurious correlations at the concept level in language models for text classification. **arXiv preprint arXiv:2311.08648**, 2023. Citado 2 vezes nas páginas 20 e 54.