UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS QUIXADÁ

CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

VICTOR EMANUEL DE SOUSA COSTA

IMPACTO DO PRÉ-TREINAMENTO VIA MLM NO DESEMPENHO E NA INTERPRETABILIDADE DO BERTIMBAU PARA A CLASSIFICAÇÃO DE NOTÍCIAS FALSAS

QUIXADÁ

2026

VICTOR EMANUEL DE SOUSA COSTA

IMPACTO DO PRÉ-TREINAMENTO VIA MLM NO DESEMPENHO E NA INTERPRETABILIDADE DO BERTIMBAU PARA A CLASSIFICAÇÃO DE NOTÍCIAS FALSAS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Orientadora: Prof. Dra. Lívia Almada Cruz.
Coorientador: Prof. Me. Décio Gonçalves de Aguiar Neto.

QUIXADÁ
2026

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

C876i Costa, Victor Emanuel de Sousa.
Impacto do Pré-treinamento via MLM no Desempenho e na Interpretabilidade do BERTimbau para a Classificação de Notícias Falsas / Victor Emanuel de Sousa Costa. – 2026.
82 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, , Quixadá, 2026.
Orientação: Profa. Dra. Lívia Almada Cruz.
Coorientação: Prof. Me. Décio Gonçalves de Aguiar Neto.

1. Notícias falsas. 2. BERTimbau. 3. XAI. 4. LIME. 5. Deriva de conceito. I. Título.
CDD

VICTOR EMANUEL DE SOUSA COSTA

IMPACTO DO PRÉ-TREINAMENTO VIA MLM NO DESEMPENHO E NA INTERPRETABILIDADE DO BERTIMBAU PARA A CLASSIFICAÇÃO DE NOTÍCIAS FALSAS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Aprovada em: 26/01/2026.

BANCA EXAMINADORA

Prof. Dra. Lívia Almada Cruz (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Me. Décio Gonçalves de Aguiar Neto (Coorientador)
Centro Universitário Católica de Quixadá (UniCatólica)

Prof. Dr. Régis Pires Magalhães
Universidade Federal do Ceará (UFC)

Prof. Dr. Paulo de Tarso Guerra Oliveira
Universidade Federal do Ceará (UFC)

Dedico este trabalho à minha mãe, ao meu pai, aos meus avós e a todos que me apoiaram ao longo desta jornada.

AGRADECIMENTOS

Agradeço, com reverência e devoção, a Deus, por ter conduzido meus passos ao longo desta jornada.

Agradeço aos meus pais, Vanessa e Wilson, que sempre fizeram de tudo para que eu vivesse bem e tivesse uma educação de qualidade. Nada do que hoje sou ou alcancei teria sido possível sem o amor e o sacrifício de cada um.

Aos meus avós, Vitorino e Clara, o meu mais sincero e profundo obrigado por serem pilares na minha vida e por sempre me ensinarem o caminho da retidão. Aos demais familiares que, de alguma forma, contribuíram para o meu futuro, agradeço imensamente a todos.

Aos amigos feitos durante essa caminhada, Deivid, Erick, Guilherme, Gustavo, Jeferson, Joabe, Jorge, João Pedro, Kaynan, Mário, Matheus, Venicius e Iarley, obrigado por todo o companheirismo e apoio, por terem dividido o peso da bagagem comigo e por garantirem que nenhum de nós ficasse pelo caminho. Vocês foram fundamentais para tornar essa trajetória mais leve.

À Ordem DeMolay, que me preparou para os deveres da vida adulta, e aos meus irmãos, agradeço por cada ensinamento transmitido, por cada palavra de incentivo e por me mostrarem o verdadeiro valor do companheirismo.

Agradeço à minha orientadora, Dra. Lívia Almada Cruz, e ao meu coorientador, Me. Décio Gonçalves de Aguiar Neto, pela excelente orientação e por toda a paciência que tiveram comigo ao longo deste processo.

Aos professores participantes da banca examinadora, pelos quais tenho grande apreço, Dr. Régis Pires Magalhães e Paulo de Tarso Guerra Oliveira, agradeço pelo tempo dedicado, bem como pelas valiosas contribuições e sugestões.

À Universidade Federal do Ceará - Campus Quixadá, e a todo o corpo docente do curso de Ciência da Computação, agradeço por todo o conhecimento acadêmico e pessoal proporcionado ao longo da graduação.

Encerrar esta etapa é, ao mesmo tempo, gratificante e comovente. A todos que, de alguma forma, contribuíram para que esta graduação se concretizasse, deixo minha sincera e eterna gratidão.

"Tudo o que pedires em oração, crendo,
recebereis" (Mateus 21:22)

RESUMO

O aumento da desinformação nas redes sociais e plataformas digitais representa um dos maiores desafios da era da informação, impactando a formação da opinião pública e ameaçando a confiança em instituições e processos democráticos. Notícias falsas têm se espalhado com rapidez e alcance inéditos, impulsionadas pela dinâmica dos algoritmos e pelo consumo acelerado de conteúdos online. Diante desse cenário, este trabalho investigou o impacto do pré-treinamento via Modelo de Linguagem Mascarado (*Masked Language Model*) no desempenho e na interpretabilidade do modelo BERTimbau para a classificação de notícias falsas em português. Utilizando o *corpus* FakeRecogna 2.0 e uma metodologia baseada em validação cruzada *StratifiedGroupK-Fold* com 10 *folds*, comparou-se um modelo de referência a uma versão especializada no domínio jornalístico brasileiro. A aplicação da técnica de Inteligência Artificial Explicável, o LIME, permitiu auditar o comportamento do classificador frente ao fenômeno de deriva de conceito em três janelas temporais, marcos estes previamente identificados na literatura como pontos de mudança significativa para este conjunto de dados. Os resultados quantitativos indicaram um ganho de desempenho com o uso do pré-treinamento via Modelo de Linguagem Mascarado, com o *F1-Score* médio elevando-se de 0,8936 para 0,9708 e a revocação média saltando de 0,8250 para 0,9479, sugerindo que a adaptação ao domínio pode ter ampliado a sensibilidade do modelo a nuances textuais que anteriormente resultavam em falsos negativos. Sob a ótica qualitativa, a análise levanta a hipótese de que o classificador correlaciona a coesão sintática e a frequência de *stopwords* com a veracidade jornalística, ao passo que parece utilizar padrões estruturais de viralização como indicativos de conteúdos enganosos. Observou-se, ainda, uma tendência de consistência temática sanitária nas notícias verdadeiras ao longo dos três períodos, enquanto as falsas aparentaram migrar mais rapidamente para a polarização política. Apesar dos métricas elevadas, notaram-se limitações na interpretação de metáforas e possíveis ruídos decorrentes de metadados de agências de checagem, reforçando que a especialização de domínio e a explicabilidade são ferramentas relevantes para a busca por robustez e transparência na classificação de notícias falsas em português.

Palavras-chave: Notícias falsas; BERTimbau; XAI; LIME; Deriva de conceito.

ABSTRACT

The surge of misinformation on social media and digital platforms constitutes one of the major challenges of the information age, impacting public opinion formation and threatening trust in democratic institutions and processes. Fake news has spread with unprecedented speed and reach, driven by algorithmic dynamics and accelerated online content consumption. Given this scenario, this work investigated the impact of Masked Language Model pre-training on the performance and interpretability of the BERTimbau model for fake news classification in Portuguese. Using the FakeRecogna 2.0 corpus and a methodology based on StratifiedGroupKFold cross-validation with 10 folds, a reference model was compared to a version specialized in the Brazilian journalistic domain. The application of the eXplainable Artificial Intelligence technique, LIME, allowed auditing the classifier's behavior regarding the concept drift phenomenon across three temporal windows, landmarks previously identified in the literature as significant change points for this dataset. Quantitative results indicated a performance gain using MLM pre-training, with the average F1-Score rising from 0.8936 to 0.9708 and average recall jumping from 0.8250 to 0.9479, suggesting that domain adaptation may have enhanced the model's sensitivity to textual nuances that previously resulted in false negatives. From a qualitative perspective, the analysis raises the hypothesis that the classifier correlates syntactic cohesion and stopword frequency with journalistic veracity, whereas it appears to use structural viralization patterns as indicators of deceptive content. Furthermore, a trend of sanitary thematic consistency was observed in true news throughout the three periods, while fake news appeared to migrate more rapidly towards political polarization. Despite high metrics, limitations were noted in interpreting metaphors and potential noise stemming from fact-checking agency metadata, reinforcing that domain specialization and explainability are relevant tools in the pursuit of robustness and transparency in fake news classification in Portuguese.

Keywords: Fake news; BERTimbau; XAI; LIME; Concept drift.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
AP	Aprendizado Profundo
BERT	*Bidirectional Encoder Representations from Transformers*
FN	Falsos Negativos
FP	Falsos Positivos
IA	Inteligência Artificial
LIME	*Local Interpretable Model-Agnostic Explanations*
MLM	*Masked Language Model*
NSP	*Next Sentence Prediction*
PLN	Processamento de Linguagem Natural
RNA	Rede Neural Artificial
RNN	*Recurrent Neural Networks*
SHAP	*SHapley Additive exPlanations*
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos
XAI	*eXplainable Artificial Intelligence*

SUMÁRIO

1 INTRODUÇÃO

O aumento da desinformação nas redes sociais e plataformas digitais representa um dos desafios mais significativos da era da informação, impactando diretamente a formação da opinião pública e a dinâmica social global. Esse fenômeno, que se intensificou nos últimos anos, ganhou proporções globais com o crescimento explosivo da circulação de notícias falsas (Shu *et al.*, 2020). Com o advento dessas plataformas, a velocidade de disseminação de informações nunca foi tão alta, permitindo que conteúdos falsos ou distorcidos alcancem milhões de usuários em questão de minutos. Estudos demonstram que informações falsas se espalham mais rápido e alcançam públicos maiores do que as verdadeiras, especialmente no contexto político (Vosoughi *et al.*, 2018).

Diante dessa realidade, pesquisadores têm recorrido a abordagens de Inteligência Artificial (IA) como uma alternativa eficaz para a classificação automatizada de notícias falsas, especialmente por meio de técnicas de Aprendizado de Máquina (AM) e Processamento de Linguagem Natural (PLN) (Shu *et al.*, 2020). Com o avanço dessas áreas, observa-se uma crescente adoção de modelos de linguagem pré-treinados baseados na arquitetura *Transformer*, como o *Bidirectional Encoder Representations from Transformers* (BERT), que alcançaram o estado da arte em diversas tarefas textuais.

No entanto, um desafio crítico na detecção de desinformação é a natureza dinâmica do conteúdo: os tópicos, o vocabulário e as estratégias de escrita mudam rapidamente ao longo do tempo (fenômeno conhecido como deriva de conceito) (Wanderley *et al.*, 2025). Modelos treinados em dados estáticos podem sofrer degradação de desempenho ou aprender atalhos linguísticos espúrios que não se sustentam no futuro. Além disso, redes neurais profundas operam frequentemente como "caixas-pretas", dificultando o entendimento dos critérios que levam às suas decisões (Silva *et al.*, 2024). Essa opacidade compromete a confiabilidade dos sistemas e levanta questionamentos éticos sobre o uso de IA em contextos sensíveis, como o jornalismo e a justiça (Alves; Andrade, 2022).

Nesse cenário, a área de *eXplainable Artificial Intelligence* (XAI) surge como uma resposta necessária. O objetivo da XAI é tornar os sistemas mais transparentes, explicando suas decisões de forma compreensível para humanos. Mais do que apenas justificar uma predição, técnicas de explicabilidade local, como o *Local Interpretable Model-Agnostic Explanations* (LIME), podem atuar como ferramentas de diagnóstico, permitindo investigar se o modelo está baseando suas decisões em características semânticas legítimas ou em vieses e ruídos do conjunto

de dados (Molnar, 2025).

Este trabalho parte do argumento de que a aplicação de técnicas de explicabilidade em modelos de classificação de notícias falsas pode não apenas aumentar a confiança nos sistemas, mas também contribuir para o desenvolvimento de soluções mais éticas e seguras. Considerando que ainda são escassos os estudos voltados especificamente para o idioma português, mesmo diante do alto volume de desinformação presente nesse contexto linguístico, acredita-se que a integração entre IA e XAI representa uma contribuição significativa e transformadora para a área (Shu *et al.*, 2020).

Este trabalho investiga a hipótese de que a adaptação de modelos de linguagem ao domínio jornalístico de notícias em português, por meio de técnicas como o pré-treinamento via *Masked Language Model* (MLM), pode influenciar não apenas o desempenho quantitativo, mas também a robustez dos critérios de decisão do modelo ao longo do tempo. Considerando a escassez de estudos que analisem a intersecção entre BERTimbau, pré-treinamento via MLM e explicabilidade temporal, esta pesquisa busca preencher essa lacuna.

Para tanto, o presente trabalho explora a aplicação do modelo BERTimbau no *corpus* FakeRecogna 2.0, comparando uma versão de referência com uma versão submetida ao pré-treinamento via MLM. A análise utiliza o LIME para mapear a evolução dos padrões linguísticos e identificar padrões de decisão adotadas pelos classificadores, visando compreender como pré-treinamento via MLM impacta a interpretação de notícias falsas em diferentes janelas temporais.

1.1 Objetivos

Nesta seção, são apresentados o objetivo geral e os objetivos específicos deste trabalho.

1.1.1 Objetivo Geral

Avaliar o impacto do pré-treinamento via MLM no desempenho e na interpretabilidade do modelo BERTimbau para a classificação de notícias falsas, utilizando a técnica LIME para investigar investigar a robustez do classificador e a evolução dos critérios de decisão frente ao fenômeno de deriva de conceito.

1.1.2 Objetivos Específicos

- Comparar o desempenho quantitativo de um modelo BERTimbau de referência (sem pré-treinamento adicional) com uma versão submetida ao pré-treinamento via MLM no corpus FakeRecogna 2.0, utilizando métricas de acurácia, precisão, revocação e F1-Score para verificar se a técnica proporciona ganhos de eficiência na classificação.
- Aplicar a técnica LIME para gerar explicações locais das decisões de ambos os modelos (referência e pré-treinado) em amostras de diferentes períodos temporais, investigando possíveis alterações nos padrões linguísticos e no vocabulário predominante ao longo dos intervalos analisados.
- Analisar as explicações geradas para identificar padrões linguísticos e discutir como o pré-treinamento altera a sensibilidade do modelo a vieses e características do texto.

1.2 Organização do texto

O restante deste texto está organizado como descrito a seguir. No Capítulo 2, são apresentados os fundamentos teóricos e conceitos que sustentam as abordagens propostas neste trabalho. No Capítulo 3, são abordados os trabalhos relacionados sobre o uso de métodos de XAI na classificação de notícias falsas. No Capítulo 4, são descritos detalhadamente os passos metodológicos deste trabalho. Posteriormente, no Capítulo 5, são apresentados os resultados obtidos em nossos experimentos, abrangendo tanto as métricas de classificação quanto a análise de interpretabilidade. Por fim, o Capítulo 6 discute as conclusões do estudo, destacando suas principais contribuições, limitações e sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são apresentados os conceitos fundamentais para o desenvolvimento deste trabalho.

2.1 Notícias falsas

A proliferação da desinformação e da informação incorreta representa um desafio significativo para a confiança da sociedade e a integridade do discurso público na era digital (Thakar; Bhatt, 2024). Nesse contexto, o termo "notícia falsa" emergiu como um problema global (Kim *et al.*, 2021). Embora não haja uma definição universalmente aceita que englobe todas as formas de falsidade, como sátira ou conteúdo fabricado, o conceito continua a evoluir com o tempo e varia conforme o foco de diferentes pesquisas (Kim *et al.*, 2021).

Uma definição comum caracteriza notícia falsa como informação fabricada que mimetiza a forma da mídia de notícias, mas que não segue os processos editoriais nem a intenção dos meios de comunicação tradicionais (Lazer *et al.*, 2018). Algumas abordagens a definem estritamente com base na intenção de enganar e na falsidade verificável da informação (Mendonça *et al.*, 2022). Por exemplo, Allcott e Gentzkow definem notícias falsas como "artigos de notícias que são intencionalmente e verificavelmente falsos e que poderiam enganar os leitores" (Kim *et al.*, 2021). Assim, notícias falsas podem ser consideradas uma subcategoria da *misinformation*, referindo-se especificamente a informações falsas que têm o potencial de iludir ou induzir dúvidas sobre a verdade (Lazer *et al.*, 2018; Mendonça *et al.*, 2022). Além disso, essas informações podem minar a credibilidade dos veículos de notícias convencionais (Hamed *et al.*, 2023).

O ambiente digital ampliou esse fenômeno, ao permitir que o público selecione e consuma conteúdos alinhados a seus vieses cognitivos pessoais (Mendonça *et al.*, 2022), facilitando que criadores de notícias falsas atinjam seus objetivos específicos. Com o desenvolvimento das mídias sociais, as fronteiras entre criadores e consumidores de notícias tornaram-se cada vez mais tênues (Kim *et al.*, 2021). A ascensão de tecnologias de inteligência artificial também impulsionou esse processo, possibilitando o uso de ferramentas como bots sociais e mídias falsas, incluindo *deepfakes*, para gerar e distribuir desinformação com maior sofisticação e alcance (Shu *et al.*, 2017).

A disseminação de notícias falsas impacta negativamente a política, a economia e a cultura, contribuindo para a instabilidade social (Kim *et al.*, 2021). Durante períodos de crise,

como epidemias, a propagação de rumores e dicas médicas falsas se intensifica nas redes sociais, explorando o medo e a ansiedade das pessoas na ausência de informações verificadas (Hamed *et al.*, 2023). A pandemia da COVID-19 é um exemplo claro: a circulação de informações falsas afetou a saúde mental coletiva e provocou estados de ansiedade na população (Kim *et al.*, 2021). Além disso, estudos mostram que a desinformação tende a se espalhar mais rapidamente do que informações verdadeiras, impulsionada por fatores como engajamento emocional e os próprios algoritmos das plataformas digitais (Mouratidis *et al.*, 2025). Esse cenário representa um risco crescente para o bem-estar social, a confiança pública e a governança democrática.

2.2 Aprendizado de Máquina

O AM é um dos pilares da Inteligência Artificial. Seu objetivo é desenvolver algoritmos capazes de identificar padrões e tomar decisões com base em dados, sem a necessidade de instruções explícitas para cada tarefa (Cerri, 2017). Esses sistemas aprendem por meio da experiência, ajustando seus comportamentos conforme a exposição a novos dados, o que os torna extremamente úteis em domínios onde soluções tradicionais baseadas em regras não são viáveis (Kühl *et al.*, 2022).

Em essência, o AM busca automatizar processos de inferência e generalização, permitindo que máquinas adquiram conhecimento a partir de exemplos. Essa abordagem é amplamente utilizada em tarefas como reconhecimento de padrões, predição de tendências, recomendação de conteúdo e, mais recentemente, na classificação automatizada de notícias falsas (Reis *et al.*, 2019).

O desenvolvimento de soluções baseadas em AM envolve etapas fundamentais: coleta e análise exploratória dos dados, pré-processamento, seleção de atributos relevantes, particionamento em conjuntos de treinamento e teste, escolha do algoritmo apropriado e avaliação dos modelos por meio de métricas como acurácia, precisão, revocação e F1-score (Reis *et al.*, 2019).

O aprendizado supervisionado é uma das categorias de algoritmos no Aprendizado de Máquina. Nesse modelo, os algoritmos são treinados com dados de entrada acompanhados dos rótulos correspondentes, ou seja, das soluções esperadas para cada exemplo apresentado. O objetivo é permitir que o sistema aprenda a mapear as entradas para as saídas corretas, de modo a generalizar esse conhecimento para novas instâncias não vistas previamente (Géron, 2019).

Esse tipo de abordagem é empregado em tarefas de classificação, como a filtragem

automática de spam, e em problemas de regressão, como a predição de preços de imóveis com base em suas características. A classificação é uma tarefa em que o modelo atribui uma categoria ou classe a cada dado, enquanto a regressão visa prever um valor numérico contínuo. Ambas as tarefas são fundamentais para diversas aplicações práticas, desde o reconhecimento de padrões até a previsão de resultados (Géron, 2019).

Nesse contexto, o problema central deste trabalho, a classificação de notícias falsas, enquadra-se precisamente na categoria de aprendizado supervisionado. A abordagem é supervisionada porque os modelos de linguagem são treinados com um conjunto de dados em que cada notícia já possui um rótulo predefinido (ex: “falsa” ou “verdadeira”). A tarefa é de classificação, pois o objetivo é que o modelo aprenda a atribuir uma dessas categorias discretas a novas notícias, não vistas durante o treinamento, automatizando assim a identificação de desinformação.

2.3 Processamento de Linguagem Natural

O PLN é uma subárea da IA que tem como objetivo capacitar máquinas a compreender, interpretar, gerar e interagir com a linguagem humana de forma eficiente (Caseli; Nunes, 2024). O PLN busca reduzir a distância entre a comunicação natural dos seres humanos e a linguagem formal utilizada pelos sistemas computacionais, permitindo que algoritmos processem grandes volumes de texto de maneira automatizada e contextualizada (Cambria; White, 2014).

Para alcançar seus objetivos, o PLN emprega técnicas que transformam o texto, uma forma de dado não estruturado, em representações numéricas que podem ser processadas por algoritmos de aprendizado de máquina. Abordagens modernas, especialmente com o advento de modelos de linguagem profundos, focam na criação de representações contextuais ricas, conhecidas como *embeddings*, que permitem uma compreensão muito mais nuançada do significado das palavras e sentenças (Caseli; Nunes, 2024).

No contexto da classificação de notícias falsas, o PLN é crucial para a análise do conteúdo textual das notícias, possibilitando a identificação de padrões linguísticos típicos da desinformação, como o uso de linguagem sensacionalista, contradições factuais ou ausência de fontes confiáveis (Shu *et al.*, 2017). O PLN também permite uma análise semântica e pragmática mais profunda, que é essencial para identificar o contexto e as intenções subjacentes à mensagem, como manipulação de informações ou distorções do discurso. Além disso, técnicas avançadas de modelagem semântica, como os *embeddings* de palavras e as representações contextuais, possibilitam a captura de nuances complexas de significado, essenciais para discernir verdades

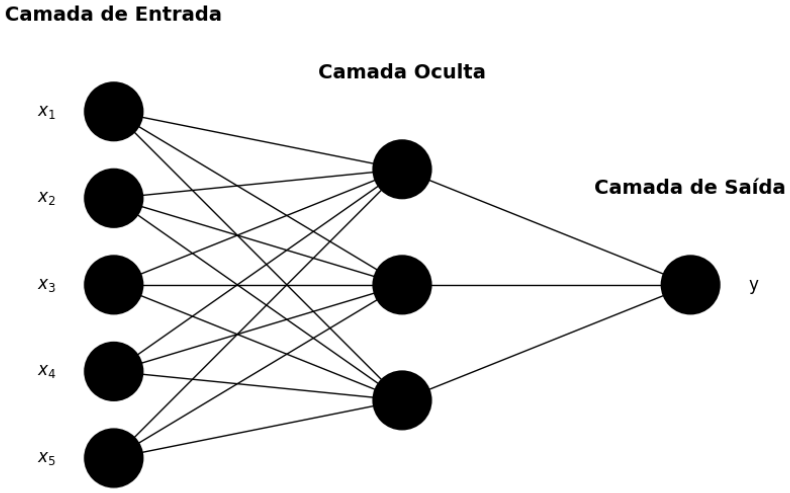
parciais ou enganosas em textos (Caseli; Nunes, 2024).

2.4 Redes Neurais Artificiais e Aprendizado Profundo

A Rede Neural Artificial (RNA) é um modelo computacional inspirado no funcionamento do cérebro biológico, projetado para aprender a partir de dados e realizar tarefas complexas como reconhecimento de padrões e classificação. (Fleck *et al.*, 2016).

Uma RNA é fundamentalmente composta por neurônios (ou nós) interconectados, onde cada conexão possui um peso sináptico que pondera a importância da entrada (Haykin, 2001). A saída de um neurônio é determinada pela combinação linear de suas entradas e pesos, seguida pela aplicação de uma função de ativação, que introduz não-linearidade e permite o modelamento de relações complexas nos dados (Rauber, 2005). A Figura 1 ilustra um exemplo de uma rede neural clássica com três camadas: entrada, oculta e saída, demonstrando o fluxo de informação *feedforward*. O conhecimento é adquirido e armazenado nesses pesos adaptáveis durante um processo iterativo de treinamento. Este treinamento pode ser supervisionado, onde o modelo aprende a partir de exemplos com saídas desejadas, ou não supervisionado, identificando padrões nos dados sem rótulos pré-definidos (Choi *et al.*, 2020).

Figura 1 – Exemplo de uma Rede Neural Clássica de 3 Camadas



Fonte: Adaptado de Tan *et al.* (2005)

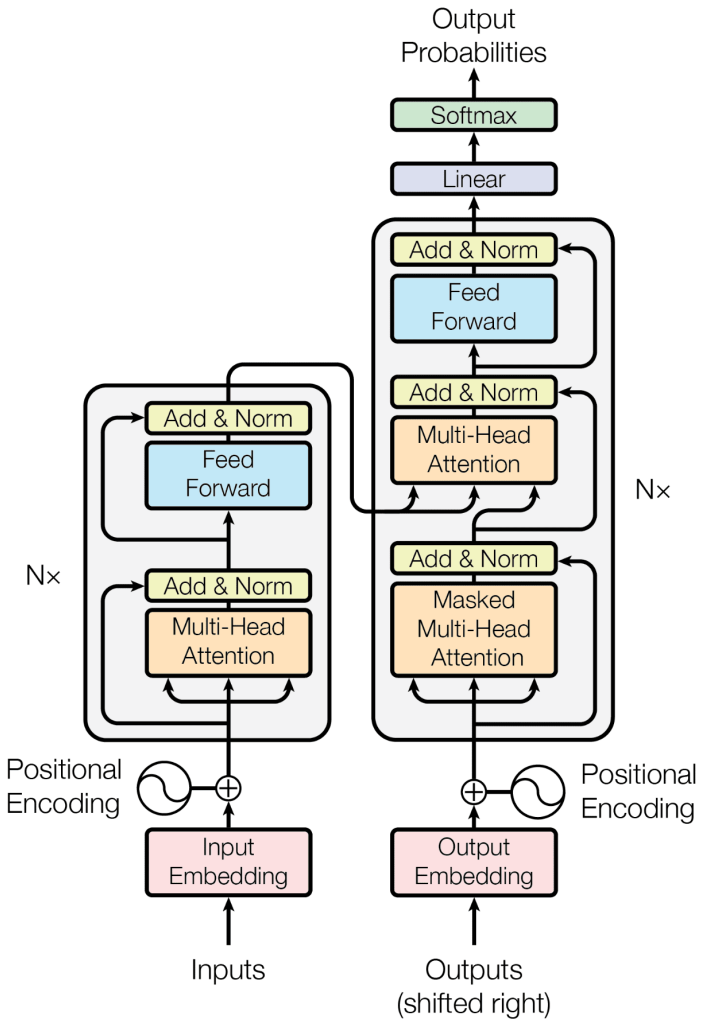
O Aprendizado Profundo (AP) emerge como um subcampo do aprendizado de máquina que corresponde a redes neurais artificiais com múltiplas camadas escondidas, que podem variar de dezenas a centenas (Chollet, 2021). A recente explosão na capacidade computacional

tem sido um fator determinante para a viabilização e ampla aplicação dos modelos de AP em diversos domínios (Fleck *et al.*, 2016).

2.4.1 Transformers

A arquitetura *Transformer*, introduzida no trabalho seminal "*Attention Is All You Need*" por Vaswani *et al.* (2017), foi criada para eliminar a necessidade de recorrência, baseando-se inteiramente em mecanismos de atenção para processar sequências. A Figura 2 ilustra a arquitetura geral de um Transformer, composta por módulos de *encoder* e *decoder*.

Figura 2 – Arquitetura *Transformer*.



Fonte: (Vaswani *et al.*, 2017)

Essencialmente, o *Transformer* é dividido em duas partes principais: o *Encoder* e o *Decoder* (Vaswani *et al.*, 2017). O *Encoder* é responsável por processar a sequência de entrada e transformá-la em uma representação rica em contexto. Ele é composto por múltiplas camadas

idênticas, cada uma contendo um mecanismo de *Multi-Head Attention* (atenção multi-cabeça) e uma rede *Feed-Forward* totalmente conectada (Vaswani *et al.*, 2017).

O mecanismo de *Multi-Head Attention* permite que o modelo pondere a importância de diferentes partes da sequência de entrada entre si, capturando relações de longa distância de forma eficiente (Vaswani *et al.*, 2017).

Uma *Positional Encoding* é adicionada às *embeddings* de entrada para fornecer informações sobre a posição dos *tokens* na sequência, já que o modelo não possui recorrência ou convolução para inferir a ordem (Vaswani *et al.*, 2017).

O *Decoder*, por sua vez, é encarregado de gerar a sequência de saída, um *token* por vez, com base na representação produzida pelo *encoder* (Vaswani *et al.*, 2017). Ele também é composto por múltiplas camadas idênticas, que incluem os dois subcomponentes presentes no *encoder* (*Multi-Head Attention* e *Feed-Forward*), e um terceiro subcomponente adicional: um mecanismo de atenção que foca na saída do *encoder* (Vaswani *et al.*, 2017).

O principal diferencial da arquitetura *Transformer* é o seu mecanismo de autoatenção (*Self-Attention*) (Vaswani *et al.*, 2017). A autoatenção permite que o modelo pondere a importância de diferentes partes da sequência de entrada ao processar cada elemento, independentemente de sua ordem, capturando relações globais entre os elementos dos dados sequenciais (Rahmadhani *et al.*, 2024). Isso contrasta com abordagens baseadas em *Recurrent Neural Networks* (RNN), que processam dados sequenciais de forma mais limitada (Tunstall *et al.*, 2022).

2.4.2 BERT

O BERT, introduzido por Devlin *et al.* (2019), revolucionou o campo do Processamento de Linguagem Natural ao propor uma metodologia de pré-treinamento baseada na arquitetura *Transformer*. Diferente de modelos anteriores que processavam texto sequencialmente, o BERT é treinado para entender o contexto de uma palavra considerando simultaneamente todos os outros termos na sentença.

Essa capacidade é alcançada por meio de uma estratégia de aprendizado multitarefa durante o pré-treinamento, fundamentada em duas técnicas principais. A primeira é o MLM, na qual o modelo deve prever palavras que foram intencionalmente mascaradas em uma sentença, baseando-se exclusivamente no contexto bidirecional fornecido pelos termos adjacentes. Em conjunto, aplica-se a técnica de *Next Sentence Prediction* (NSP), onde o objetivo é determinar se duas sentenças aparecem consecutivamente no *corpus* de treinamento, auxiliando o modelo a

compreender relações de coerência e continuidade entre frases (Devlin *et al.*, 2019).

Essa abordagem permite ao BERT gerar representações contextuais ricas, ajustáveis para diversas tarefas, como classificação de texto, análise de sentimentos e resposta a perguntas, mantendo um bom desempenho (Devlin *et al.*, 2019).

2.4.2.1 BERTimbau

Para aplicações em língua portuguesa, destaca-se o BERTimbau, uma adaptação do modelo BERT original, pré-treinado especificamente para o português brasileiro. Desenvolvido por pesquisadores brasileiros, este modelo oferece um conhecimento linguístico e contextual superior em relação a modelos multilíngues ou treinados apenas em inglês quando aplicados a dados locais (Souza *et al.*, 2020).

A arquitetura do BERTimbau segue a do BERT base, mas seu diferencial reside no treinamento extensivo sobre o *corpus* BrWaC (*Brazilian Web as Corpus*), composto por um vasto volume de páginas da web em português. Uma característica crucial do BERTimbau para tarefas de classificação é a sua capacidade de lidar com acentuação e preservar a distinção entre letras maiúsculas e minúsculas. A manutenção dos diacríticos e da caixa alta/baixa é vital para a desambiguação semântica em português, o que enriquece a representação vetorial do texto.

O modelo é disponibilizado em dois tamanhos, denominados Base e Large. A versão Base é composta por 110 milhões de parâmetros. Já a versão Large conta com 330 milhões de parâmetros.

A versão Base, especificamente, oferece robustez suficiente para capturar características linguísticas complexas, como as presentes em notícias falsas, sem incorrer na demanda excessiva de memória e tempo de processamento exigida pela versão Large.

2.5 Deriva de Conceito em Fluxos de Informação

A eficácia da maioria dos modelos de aprendizado de máquina para classificação de notícias falsas baseia-se na premissa de treinamento estático (*offline*), o que impõe limitações significativas ao lidar com a dinâmica temporal dos ecossistemas de informação do mundo real (Wanderley *et al.*, 2025). Essa abordagem torna os classificadores vulneráveis ao fenômeno conhecido como Deriva de Conceito (*Concept Drift*), que ocorre quando as propriedades estatísticas dos dados, incluindo padrões linguísticos e tópicos, evoluem ao longo do tempo (Wanderley

et al., 2025).

No contexto específico da desinformação, ignorar a ordem cronológica dos dados mascara os efeitos dessa degradação temporal. Estudos indicam que tratar o problema como estático pode levar a uma superestimação substancial do desempenho dos modelos em cenários reais. Isso é particularmente crítico durante eventos de alto impacto, como eleições ou crises sanitárias (a exemplo da pandemia de COVID-19), onde as táticas de desinformação e os vocabulários utilizados evoluem rapidamente (Wanderley *et al.*, 2025).

2.5.1 Tipologia e Impacto na Desinformação

A deriva de conceito pode se manifestar de diversas formas, incluindo mudanças súbitas, graduais ou recorrentes nas distribuições de probabilidade dos dados (Wanderley *et al.*, 2025). No domínio de notícias falsas em português brasileiro, evidências empíricas sugerem uma distinção comportamental marcante entre as classes. As notícias verdadeiras tendem a apresentar maior estabilidade semântica, pois seguem padrões jornalísticos estabelecidos e reportam fatos consistentes (Wanderley *et al.*, 2025). Em contrapartida, as notícias falsas exibem maior variabilidade semântica e temporal, adaptando narrativas para se alinharem a eventos recentes e maximizar o engajamento, ou substituindo tópicos inteiramente quando sua relevância diminui (Wanderley *et al.*, 2025).

Essa volatilidade exige que sistemas de defesa sejam capazes não apenas de classificar, mas de adaptar-se a novos contextos semânticos. A literatura aponta que a análise de deriva em línguas com menos recursos, como o português, ainda é escassa, com a maioria dos *benchmarks* focados na língua inglesa (Wanderley *et al.*, 2025).

2.5.2 Detecção de Mudanças Distribucionais

Para identificar a ocorrência de deriva, a literatura utiliza métodos estatísticos e baseados em similaridade semântica. Testes de duas amostras, como o *Kernel Two-Sample Test* (KTS) e o *Least-Squares Density Difference* (LSDD), têm se mostrado eficazes na detecção de mudanças em representações vetoriais de texto de alta dimensão, como *embeddings* gerados pelo BERT (Wanderley *et al.*, 2025).

Complementarmente, métricas de dissimilaridade semântica, como a distância de cosseno entre centroides de janelas temporais e a *Word Mover's Distance* (WMD), permitem quantificar a magnitude dessas mudanças (Wanderley *et al.*, 2025). Análises aplicadas ao

corpus FakeRecogna 2.0 confirmaram que o período de 2020 a 2021 foi marcado por derivas significativas impulsionadas pela pandemia, com pontos de mudança abruptos detectados tanto em notícias verdadeiras quanto falsas (Wanderley *et al.*, 2025).

2.6 Inteligência Artificial Explicável

A crescente adoção de sistemas de inteligência artificial em domínios críticos tem intensificado a demanda por modelos que não apenas tomem decisões precisas, mas também expliquem de forma clara os motivos de suas predições. Em áreas como diagnóstico médico, decisões jurídicas, sistemas financeiros e, particularmente, na classificação de notícias falsas, a transparência é fundamental para garantir a confiança dos usuários e a responsabilização das decisões automatizadas (Gunning; Aha, 2019; Arrieta *et al.*, 2020).

Modelos de aprendizado profundo, como redes neurais e *Transformers*, apresentam excelentes desempenhos, mas muitas vezes são considerados "caixas-pretas" devido à dificuldade de interpretar suas decisões. Essa opacidade pode gerar resistência à adoção de tais sistemas e dificuldades em auditorias, validação e análise de vieses (Rudin, 2019).

Nesse contexto, a Inteligência Artificial Explicável torna-se uma ferramenta essencial, pois permite compreender o comportamento do modelo, diagnosticar falhas, mitigar preconceitos algorítmicos e possibilitar intervenções humanas mais seguras. Além disso, explicações claras contribuem para a formação de usuários mais críticos, especialmente no combate à desinformação, promovendo uma análise mais consciente de conteúdos detectados como falsos (Mishima; Yamana, 2022).

Diversos métodos foram desenvolvidos para tornar os modelos de IA mais interpretáveis, abrangendo desde métodos globais até abordagens locais e agnósticas ao modelo. Entre os métodos *post-hoc* mais utilizados, que explicam o modelo após o seu treinamento, destaca-se o LIME, o qual será detalhado a seguir.

2.6.1 LIME

O LIME é um método de explicabilidade *post-hoc* que se destaca por ser agnóstico ao modelo, ou seja, pode ser aplicado a qualquer modelo de aprendizado de máquina (Ribeiro *et al.*, 2016).

O principal objetivo do LIME é gerar explicações locais para predições individuais.

Ele funciona perturbando ligeiramente a instância de entrada (por exemplo, removendo ou adicionando palavras em um texto), gerando um novo conjunto de dados sintéticos próximos à instância original (Ribeiro *et al.*, 2016).

Em seguida, um modelo de interpretabilidade simples, como uma regressão linear, é treinado nesse novo conjunto de dados, ponderando as amostras sintéticas de acordo com sua proximidade à instância original. A explicação local resultante, que é a interpretação do modelo simples, é então usada para aproximar o comportamento do modelo complexo para aquela predição específica (Ribeiro *et al.*, 2016).

Em sua implementação, o LIME frequentemente utiliza modelos lineares esparsos como explicações, que são treinados na vizinhança local da predição a ser explicada. As amostras geradas são ponderadas de acordo com sua proximidade à instância original, dando mais peso às amostras mais próximas para garantir a fidelidade local (Ribeiro *et al.*, 2016).

Essa abordagem permite que o LIME adapte sua explicação à forma como o modelo se comporta especificamente naquela região do espaço de características, sem tentar aproximar o modelo globalmente, o que seria impraticável para modelos complexos e não lineares (Ribeiro *et al.*, 2016).

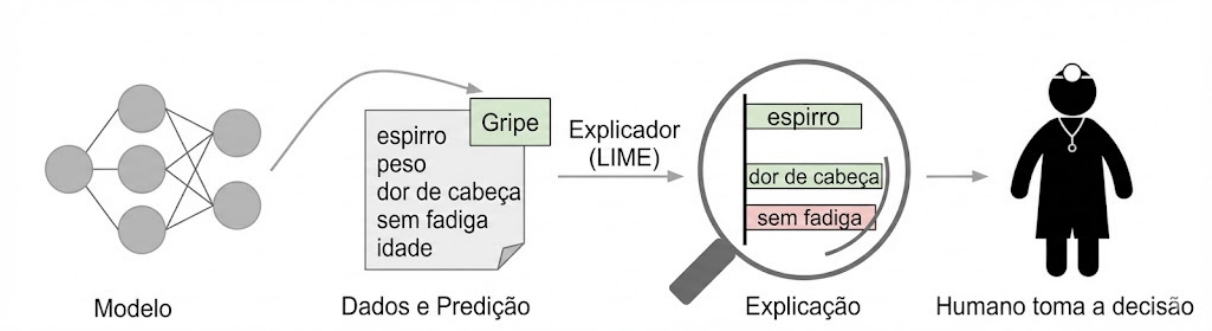
Em termos de arquitetura de software, a biblioteca disponibiliza diferentes classes de explicadores otimizados para domínios específicos, como o `LimeTabularExplainer` para dados estruturados e o `LimeImageExplainer` para visão computacional. Considerando a natureza textual do conjunto de dados utilizado nesta pesquisa, foi adotado o `LimeTextExplainer`.

Para viabilizar a integração entre o `LimeTextExplainer` e a arquitetura BERTimbau, foi necessário implementar uma função (*wrapper*). Essa necessidade surge devido à incompatibilidade de interfaces: o LIME opera gerando perturbações no texto bruto, enquanto o modelo BERT requer como entrada tensores numéricos processados por um tokenizador. O *wrapper* desenvolvido atua convertendo as *strings* perturbadas em tensores compatíveis com o BERT através de tokenização com truncamento e preenchimento, submetendo esses tensores ao modelo em modo de avaliação e, finalmente, aplicando a função *Softmax* aos *logits* resultantes para transformar as saídas brutas em uma distribuição de probabilidades interpretável pelo explicador.

Para concretizar o entendimento do método, a Figura 3 ilustra o fluxo de explicação de uma predição individual, em um cenário adaptado de Ribeiro *et al.* (2016).

No exemplo apresentado, um modelo prevê que um paciente está com gripe baseando-se em seus dados clínicos. O LIME atua como um explicador, destacando quais sintomas no

Figura 3 – Explicação de uma predição individual com LIME



Fonte: Traduzido de Ribeiro *et al.* (2016).

histórico do paciente conduziram a essa conclusão. As características "espirro" e "dor de cabeça" são apresentadas com barras verdes, indicando que contribuíram positivamente para a predição da classe "Gripe". Em contrapartida, a característica "sem fadiga" é destacada em vermelho, sinalizando uma evidência contrária à predição.

Com o auxílio dessa explicação visual, um médico que tomará a decisão encontra-se em uma posição muito mais favorável para avaliar a confiabilidade do sistema. Ao confrontar os pesos atribuídos pelo modelo com seu próprio conhecimento de domínio, o especialista pode decidir de forma informada se aceita ou rejeita a predição do modelo, mitigando riscos de diagnósticos incorretos baseados em correlações espúrias (Ribeiro *et al.*, 2016).

2.7 Métricas de Avaliação dos Modelos

Nesta seção, serão apresentadas as principais métricas utilizadas para avaliar o desempenho dos modelos empregados na classificação de notícias falsas. Essas métricas são essenciais para compreender o comportamento dos modelos de classificação e determinar sua eficácia na identificação correta das instâncias de interesse. A seguir, serão discutidas as métricas de acurácia, precisão, revocação, F1-score e, em detalhe, a matriz de confusão.

2.7.1 Matriz de Confusão

A matriz de confusão é uma ferramenta fundamental para analisar o desempenho de um modelo de classificação. Trata-se de uma tabela que sumariza os resultados de uma classificação, exibindo os números de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos. Essa estrutura permite uma análise detalhada dos tipos de acertos e erros cometidos pelo modelo (Géron, 2019).

Em problemas de classificação binária, os termos da matriz de confusão são empre-

gados para avaliar o desempenho do modelo em relação às classes positiva e negativa. Cada instância é categorizada em um dos quatro tipos a seguir:

- Verdadeiros Positivos (VP): Instâncias da classe positiva que foram corretamente classificadas como positivas pelo modelo.
- Verdadeiros Negativos (VN): Instâncias da classe negativa que foram corretamente classificadas como negativas pelo modelo.
- Falsos Positivos (FP): Instâncias da classe negativa que foram incorretamente classificadas como positivas pelo modelo.
- Falsos Negativos (FN): Instâncias da classe positiva que foram incorretamente classificadas como negativas pelo modelo.

Esses quatro termos são cruciais para a compreensão aprofundada do desempenho de um modelo de classificação. Com base neles, é possível derivar métricas importantes como precisão, revocação e F1-score, as quais fornecem uma medida mais robusta da eficácia do modelo em diferenciar entre as classes (Géron, 2019).

A Figura 4 ilustra um exemplo de matriz de confusão para um problema de classificação binária, destacando a localização dos Verdadeiros Negativos, Verdadeiros Positivos, Falsos Positivos e Falsos Negativos.

2.7.2 Acurácia

A acurácia é uma das métricas mais intuitivas e amplamente utilizadas para avaliar o desempenho geral de um modelo de classificação. Ela mede a proporção de previsões corretas, tanto de notícias classificadas como falsas quanto verdadeiras, em relação ao número total de instâncias avaliadas. Em outras palavras, a acurácia indica a porcentagem de vezes que o modelo acertou sua predição.

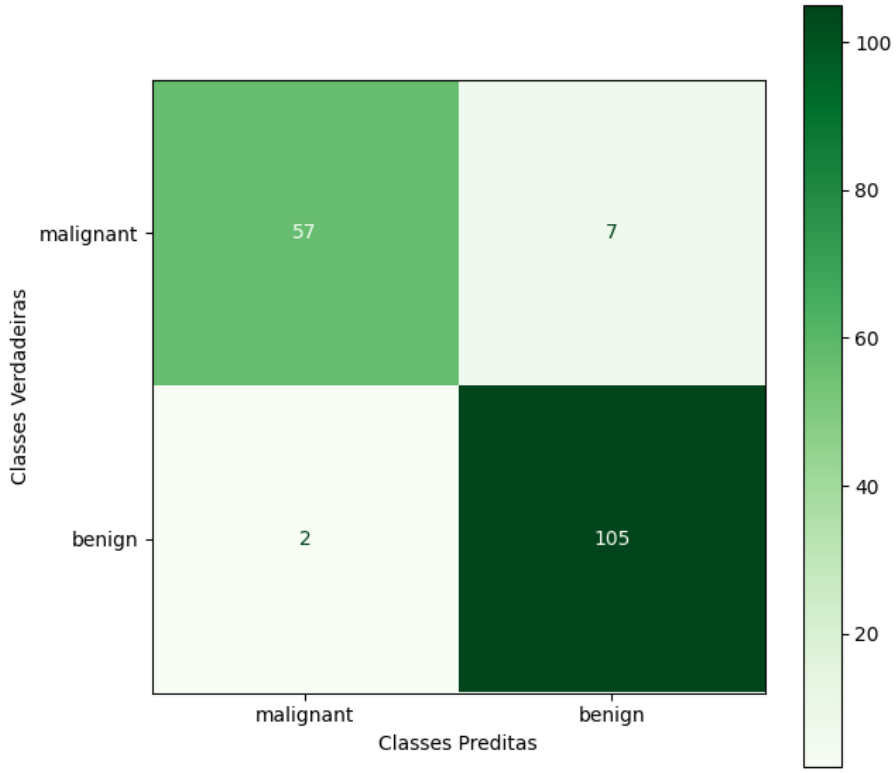
Matematicamente, a acurácia é definida pela seguinte fórmula, utilizando os termos da matriz de confusão:

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

Ou, de forma simplificada:

$$\text{Acurácia} = \frac{\text{Número de Previsões Corretas}}{\text{Número Total de Previsões}}$$

Figura 4 – Matriz de Confusão Binária de um conjunto de dados sobre câncer de mama.



Fonte: Elaborado pelo autor.

Embora a acurácia seja fácil de entender e calcular, ela pode ser uma métrica enganosa, especialmente em cenários onde as classes são desbalanceadas. No contexto da classificação de notícias falsas, por exemplo, se a maioria das notícias em um conjunto de dados é verdadeira e apenas uma pequena parcela é falsa, um modelo que simplesmente prevê "notícia verdadeira" para todas as instâncias pode alcançar uma alta acurácia (ex: 95%), sem, contudo, ser eficaz na identificação das notícias falsas, a classe minoritária e de maior interesse (Géron, 2019). Por essa razão, é crucial analisar a acurácia em conjunto com outras métricas, como precisão, revocação e F1-score, para obter uma avaliação mais completa e justa da capacidade do modelo em identificar notícias falsas.

2.7.3 Precisão

A precisão é uma métrica que avalia a qualidade das previsões positivas de um modelo. Ela responde à pergunta: "De todas as instâncias que o modelo classificou como positivas, quantas realmente eram positivas?". Em outras palavras, a precisão mede a proporção de previsões positivas corretas em relação ao total de previsões positivas realizadas pelo modelo (Géron, 2019).

No contexto da classificação de notícias falsas, a precisão indica a proporção de notícias que o modelo classificou como falsas e que de fato eram notícias falsas. Uma alta precisão significa que, quando o modelo aponta uma notícia como falsa, ele está, na maioria das vezes, correto, minimizando o número de notícias verdadeiras classificadas erroneamente como falsas. Isso é particularmente importante para evitar a desinformação de "alarmes falsos", onde notícias legítimas são marcadas como notícias falsas, potencialmente gerando desconfiança indevida.

Matematicamente, a precisão é definida como:

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

A precisão é uma métrica importante quando o custo de um falso positivo é elevado. Por exemplo, em sistemas de saúde, um falso positivo para uma doença grave pode levar a tratamentos desnecessários e ansiedade. Na classificação de notícias falsas, um falso positivo significa que uma notícia verdadeira é indevidamente rotulada como falsa, o que pode prejudicar a reputação da fonte ou do veículo de comunicação. Por isso, buscar uma alta precisão é fundamental quando o objetivo é ser muito seletivo e confiável nas classificações positivas.

2.7.4 Revocação

A revocação é uma métrica que avalia a capacidade do modelo de identificar corretamente todas as instâncias positivas reais. Ela responde à pergunta: "De todas as instâncias que realmente são positivas, quantas o modelo conseguiu identificar corretamente?". Em outras palavras, a revocação mede a proporção de verdadeiros positivos que foram encontrados pelo modelo (Géron, 2019).

No contexto da classificação de notícias falsas, a revocação é a proporção de todas as notícias falsas reais que o modelo conseguiu classificar corretamente como falsas. Uma alta revocação significa que o modelo é muito bom em "capturar" a grande maioria das notícias falsas existentes, minimizando o número de notícias falsas que passam despercebidas, os falsos negativos. Essa métrica é particularmente crítica quando o custo de um FN é alto, ou seja, quando é crucial evitar que instâncias positivas sejam classificadas erroneamente como negativas.

Matematicamente, a revocação é definida como:

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

A importância da revocação é evidente em cenários onde "perder" uma instância positiva tem consequências graves. Por exemplo, em diagnósticos médicos, um falso negativo pode significar não identificar uma doença em um paciente que realmente a possui, atrasando o tratamento. Na classificação de notícias falsas, um falso negativo implica que uma notícia falsa continua a se espalhar e potencialmente causar danos, pois o modelo falhou em identificá-la. Portanto, buscar uma alta revocação é fundamental quando o objetivo principal é minimizar os casos de notícias falsas não detectadas, mesmo que isso possa resultar em mais falsos positivos (Géron, 2019).

2.7.5 F1-Score

O F1-Score é uma métrica que busca equilibrar a precisão e a revocação de um modelo, sendo particularmente útil em cenários onde existe um desbalanceamento entre as classes ou quando ambas as métricas são importantes. Ele representa a média harmônica da precisão e da revocação, oferecendo uma pontuação única que penaliza modelos com desempenho muito baixo em qualquer uma dessas duas métricas. Um alto F1-Score indica que o modelo possui tanto alta precisão quanto alta revocação (Géron, 2019).

No contexto da classificação de notícias falsas, o F1-Score é valioso porque tanto classificar uma notícia verdadeira como falsa (falso positivo) quanto deixar uma notícia falsa passar como verdadeira (falso negativo) são erros com consequências significativas. O F1-Score fornece uma visão mais completa do desempenho do modelo ao considerar ambos os tipos de erros. Se, por exemplo, um modelo tem uma precisão altíssima mas uma revocação muito baixa (detecta pouquíssimas notícias falsas, mas as que detecta são corretas), ou vice-versa, o F1-Score será menor, indicando que o modelo não está performando bem de forma equilibrada.

Matematicamente, o F1-Score é calculado como:

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

A escolha de utilizar o F1-Score é apropriada quando o objetivo é encontrar um modelo que seja bom tanto em identificar corretamente as notícias falsas (revocação) quanto em

evitar classificar indevidamente notícias verdadeiras como falsas (*precisão*). Ele oferece uma medida mais robusta do desempenho geral em comparação com a *acurácia*, especialmente em conjuntos de dados desbalanceados (Géron, 2019).

3 TRABALHOS RELACIONADOS

Neste capítulo, são apresentados trabalhos que abordam aspectos semelhantes à proposta deste trabalho.

3.1 Comparando Técnicas de Explicabilidade sobre Modelos de Linguagem: um Estudo de Caso na Detecção de Notícias Falsas

Vicentini (2024) investigaram a aplicação de métodos de XAI para interpretar decisões de modelos de linguagem no contexto da desinformação em português. A pesquisa utilizou os conjuntos de dados FakeRecogna e Fake.Br Corpus para treinar quatro variações de modelos, visando comparar o impacto da remoção de *stopwords* no desempenho preditivo. Os resultados evidenciaram que os modelos que mantiveram as *stopwords* obtiveram os melhores desempenhos, alcançando acurácia superior a 99%, o que indica que esses termos são fundamentais para a classificação devido à sua distribuição estrutural no texto, e não apenas à sua frequência.

No âmbito da explicabilidade, o trabalho comparou os algoritmos LIME e *Integrated Gradients* (IG). A análise demonstrou que o LIME apresentou resultados mais interpretáveis, identificando palavras-chave de forma sucinta, enquanto o (IG) mostrou-se mais desafiador devido a sutilezas na visualização dos mapas de saliência. O estudo revelou padrões linguísticos distintos: notícias falsas foram frequentemente associadas a uma linguagem informal e pessoal, termos como “oi”, “pessoal”, “vocês”, e vocabulário sensacionalista, como “absurdos”, “picareta”. Em contraste, notícias reais apresentaram maior coesão, uso frequente de conectivos e citação de fontes de verificação confiáveis, como “Lupa” e “UOL”.

Outro achado relevante refere-se à variação dos pesos de entidades políticas conforme o período: no FakeRecogna (2019-2021), nomes de figuras políticas muitas vezes contribuíram para a classificação como notícia real, enquanto no Fake.Br (2016-2018), termos como “corrupção” e nomes de magistrados foram associados a conteúdos falsos. Adicionalmente, validações estatísticas com a biblioteca *spaCy* confirmaram que notícias reais são significativamente mais longas, buscando apresentar mais detalhes. Os autores concluem que, embora eficientes para níveis morfológicos e lexicais, os métodos de XAI ainda possuem limitações para capturar nuances morfossintáticas e retóricas mais profundas.

Este estudo apresenta forte convergência com a presente pesquisa ao focar no cenário brasileiro e validar a eficácia do LIME e a importância das *stopwords* na estrutura de notícias

reais. Contudo, enquanto Vicentini (2024) comparam diferentes técnicas de explicabilidade, este trabalho se aprofunda no impacto do pré-treinamento via MLM no modelo BERTimbau e investiga, de forma específica, como os padrões de decisão e vieses do modelo evoluem temporalmente dentro do mesmo domínio.

3.2 *Truth Is All You Need: Enhancing Fake News Detection with Interpretable Language Models*

Danesh e Rezanejad (2025) propõem um framework para mitigar a natureza de caixa-preta de modelos de linguagem aplicados à detecção de desinformação, utilizando o modelo BERT submetido a ajuste fino no *COVID-19 Fake News Dataset*. A análise exploratória do estudo revelou distinções estruturais marcantes: notícias verdadeiras tendem a ser consideravelmente mais extensas, possuem maior presença de URLs (69,2% contra 32,7% em notícias falsas) e citam fontes confiáveis com mais frequência do que as notícias falsas. Além disso, os autores identificaram que o vocabulário das fake news é frequentemente carregado de termos emocionais ou conspiratórios, enquanto as notícias reais focam em termos baseados em dados e relatórios oficiais.

Para garantir transparência, a abordagem integrou duas técnicas de XAI: o LIME, para quantificar a contribuição local de palavras, e a visualização dos pesos de atenção do BERT. O modelo proposto alcançou acurácia de 97,66% e F1-Score de 97,49%, superando outros os algoritmos avaliados. Comparativamente, o *Support Vector Machine* (SVM) obteve 93,32% de acurácia, seguido pela Regressão Logística com 91,96%, o *Gradient Boosted Decision Tree* (GDBT) com 86,96% e, por fim, a Árvore de Decisão com 85,37%. A análise de interpretabilidade demonstrou que o modelo associava termos como “conspiração do governo” e nomes de figuras políticas a conteúdos falsos, enquanto termos como “casos reportados” e “testagem” eram indicadores de veracidade. O estudo conclui que o conceito de “verdade” na predição de máquinas deve contemplar não apenas a precisão estatística, mas também a capacidade de explicação, sendo este um requisito essencial para a implementação de sistemas automáticos de detecção de notícias falsas em cenários do mundo real, onde a confiança e a responsabilidade ética são fundamentais.

Este trabalho correlaciona-se com esta pesquisa ao adotar a arquitetura BERT e a técnica LIME como pilares para a classificação e explicação de notícias falsas. No entanto, o presente trabalho distingue-se por investigar o cenário da língua portuguesa, aplicando o

modelo BERTimbau ao *corpus* FakeRecogna 2.0. Além disso, esta pesquisa inova ao analisar especificamente o impacto do pré-treinamento via MLM na evolução dos padrões linguísticos e nos critérios de decisão do modelo.

3.3 *A New cross-domain strategy based XAI models for fake news detection*

Kanneganti (2023) propõe uma abordagem para a detecção de notícias falsas fundamentada na integração de modelos de XAI com uma estratégia de domínio cruzado (*cross-domain*). O objetivo central da pesquisa é aplicar o aprendizado por transferência para permitir que o conhecimento extraído de um domínio de origem seja eficaz na classificação de textos em um domínio de destino, garantindo simultaneamente a compreensibilidade do modelo.

Para operacionalizar essa estratégia, o estudo utiliza o modelo BERT submetido a ajuste fino. A metodologia destaca-se pela definição de uma estratégia de quatro níveis baseada no intervalo de distribuição (*gap*) entre os dados de treino e teste: o Nível 1 representa o menor *gap* (mesmo domínio), enquanto o Nível 4 apresenta o maior desafio, onde não há relação direta entre as bases além do contexto geral de notícias.

Para prover transparência, foram integrados quatro modelos de explicabilidade local e agnósticos: LIME, que utiliza aproximações lineares locais; *Anchor*, baseado em regras de decisão do tipo "se-então"; *SHapley Additive exPlanations* (SHAP), fundamentado na teoria dos jogos para mensurar a contribuição de características; e *ELI5*, que destaca pesos de palavras. Os resultados experimentais evidenciaram a sensibilidade do BERT à variação de domínios, com a acurácia decaindo de 84% no Nível 1 para 57% no Nível 4. A análise permitiu identificar pares ideais de ferramentas XAI para diferentes contextos (ex: LIME e *Anchor* para mesmo domínio; LIME e SHAP para alto *gap*), concluindo que a explicabilidade é crucial para validar predições antes de sua utilização em tomadas de decisão governamentais ou corporativas.

A pesquisa de Kanneganti (2023) dialoga diretamente com o presente trabalho ao demonstrar a necessidade de adaptar modelos BERT para novos contextos de distribuição de dados e ao utilizar o LIME como ferramenta central de validação. No entanto, enquanto Kanneganti foca na variação temática, esta pesquisa investiga a variação temporal no cenário da língua portuguesa. Além disso, propõe-se aqui analisar especificamente como o pré-treinamento via MLM no modelo BERTimbau influencia a evolução dos padrões linguísticos e a robustez do classificador ao longo do tempo.

3.4 Análise Comparativa

O Quadro 1 apresenta uma síntese das semelhanças e diferenças entre os trabalhos relacionados descritos neste capítulo e a presente pesquisa. A análise permite situar este estudo, evidenciando as escolhas metodológicas comuns e as lacunas que esta proposta visa preencher.

Observa-se uma convergência metodológica no uso de modelos baseados na arquitetura BERT como estado da arte para a classificação de textos. Tanto Danesh e Rezanejad (2025) quanto Kanneganti (2023) utilizam o BERT padrão para o idioma inglês, enquanto Vicentini (2024) e este trabalho adotam modelos específicos para a língua portuguesa, alinhando-se à necessidade de processamento de linguagem natural focado em idiomas específicos.

No que tange à explicabilidade, o algoritmo LIME figura como uma ferramenta recorrente nos trabalhos analisados, embora não represente necessariamente o estado da arte atual para a interpretação profunda de modelos de linguagem. Enquanto Kanneganti (2023) e Vicentini (2024) confrontam o LIME com técnicas teoricamente mais robustas, como SHAP e *Integrated Gradients*, a adoção do LIME nestes estudos, assim como na presente pesquisa, fundamenta-se em sua facilidade de análise visual. A escolha desta técnica não decorre de uma superioridade arquitetural sobre métodos baseados em gradientes ou atenção (abordados por Danesh e Rezanejad (2025)), mas sim de sua aptidão para gerar explicações locais agnósticas e visualmente intuitivas. Tal característica viabiliza a inspeção humana direta de amostras isoladas, alinhando-se ao objetivo diagnóstico deste trabalho: mapear pontualmente a influência do vocabulário nas decisões do classificador ao longo das janelas temporais.

Em relação ao idioma e aos conjuntos de dados, nota-se uma divergência. A literatura internacional, representada por Danesh e Rezanejad (2025) e Kanneganti (2023), dispõe de uma variedade de conjuntos de dados temáticos (como COVID-19) ou de múltiplos domínios. No cenário brasileiro, Vicentini (2024) e a presente pesquisa concentram-se em conjuntos de dados em português. O presente estudo inova ao utilizar a versão mais recente, o FakeRecogna 2.0, e ao explorar a dimensão temporal desses dados.

O principal diferencial desta pesquisa reside no foco do problema abordado. Enquanto Vicentini (2024) confrontam técnicas de explicabilidade (LIME vs. IG) e analisam a relevância estrutural das *stopwords*, Danesh e Rezanejad (2025) priorizam a análise de características estilísticas e Kanneganti (2023) exploram a adaptação entre domínios temáticos, este trabalho investiga explicitamente o fenômeno de deriva de conceito e avalia como o pré-treinamento via MLM no modelo BERTimbau influencia a robustez e os padrões explicativos do

modelo ao longo do tempo.

Quadro 1 – Quadro Comparativo entre Trabalhos Relacionados

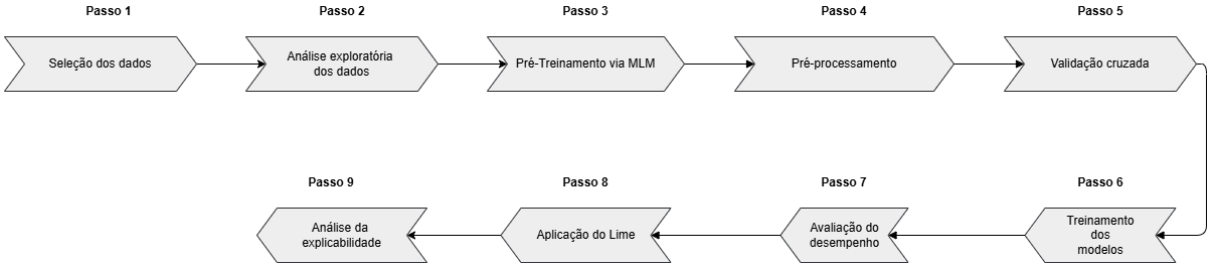
Trabalho	**Modelo utilizado**	**Técnicas de XAI**	**Conjunto de Dados**	**Idioma**	**Foco Principal**
Vicentini (2024)	BERTimbau	LIME, IG (*Integrated Gradients*)	FakeRecogna, Fake.Br *Corpus*	PT	Comparação de técnicas de XAI e impacto de *stopwords*.
Danesh e Rezanejad (2025)	BERT	LIME, *Attention Weights*	*COVID-19 Fake News Dataset*	EN	Análise estilística e estrutural em contexto de pandemia.
Kanneganti (2023)	BERT	LIME, Anchor, SHAP, ELI5	COVID-19, Mistos e *Cross-domain*	EN	Estratégia de adaptação entre domínios temáticos distintos.
Este Trabalho	BERTimbau	LIME	FakeRecogna 2.0	PT	Aplicação de explicabilidade para interpretar o impacto do pré-treinamento MLM e a evolução temporal (deriva de conceito).

Fonte: Elaborado pelo autor.

4 METODOLOGIA

Neste capítulo, são descritos os procedimentos metodológicos do trabalho, abrangendo a seleção dos dados, o pré-treinamento via MLM do modelo de linguagem, a estratégia de validação cruzada e os métodos de explicabilidade. A abordagem propõe uma etapa preliminar de pré-treinamento via MLM para especializar o modelo no contexto de notícias em português, seguida de uma classificação supervisionada. Todas as etapas descritas nesse Capítulo foram implementadas na linguagem Python, utilizando o ambiente Google Colab. A Figura 5 ilustra o fluxograma do processo de execução do trabalho.

Figura 5 – Passos de execução do trabalho



Fonte: Elaborado pelo autor.

4.1 Seleção dos dados

No contexto da classificação de notícias falsas, a escolha do conjunto de dados configura-se como uma etapa essencial, pois impacta diretamente a qualidade dos experimentos e a capacidade de generalização dos modelos. De acordo com Caseli e Nunes (2024), a construção de um bom conjunto de dados linguístico deve reunir cinco características fundamentais: consistência, variedade, representatividade, documentação detalhada e tamanho.

Embora existam diversos conjuntos de dados voltados à classificação de notícias falsas na literatura, a maioria deles encontra-se disponível apenas na língua inglesa, o que dificulta a aplicação direta em contextos da língua portuguesa. Considerando esse cenário, este trabalho adota o conjunto *FakeRecogna 2.0*, por atender aos critérios de qualidade exigidos: dados rotulados, quantidade expressiva de exemplos e equilíbrio estatístico entre as classes.

O conjunto *FakeRecogna 2.0* constitui uma versão expandida do corpus original FakeRecogna, proposto por Garcia *et al.* (2024) e desenvolvido para a tarefa de classificação de notícias falsas em língua portuguesa. Esta base compreende um total de 52.800 amostras de notícias, balanceadas entre os rótulos “falsa” e “verdadeira”, extraídas entre os anos de 2002 e

2023 (Caseli; Nunes, 2024).

As notícias falsas foram coletadas de nove agências brasileiras de checagem de fatos, como Aos Fatos, Agência Lupa, Boatos.org, Estadão Verifica, Fato ou Fake, Projeto Comprova, entre outras. As notícias verdadeiras, por sua vez, foram extraídas de portais jornalísticos confiáveis, como G1, UOL, Extra e da página oficial do Ministério da Saúde. A coleta foi realizada por meio de técnicas de *web crawling*, especificamente projetadas para acessar fontes relevantes e garantir diversidade temática (Garcia *et al.*, 2024).

Além do conteúdo textual principal, cada instância do conjunto contém metadados adicionais relevantes, como título, subtítulo, categoria, data de publicação, autor e URL original da notícia. Para o desenvolvimento deste trabalho, utilizou-se a versão do conjunto de dados disponibilizada publicamente pelos autores no repositório da plataforma Hugging Face[1], o que favorece a reprodutibilidade dos experimentos.

4.2 Análise Exploratória dos Dados

Antes de submeter os dados aos modelos, conduziu-se uma análise exploratória visando assegurar a integridade estrutural do conjunto de dados e verificar as premissas de balanceamento necessárias para o treinamento supervisionado.

O procedimento analítico consistiu na contagem das amostras válidas remanescentes, segmentadas de acordo com os rótulos presentes na coluna *"Label"*. Essa contabilização teve como objetivo diagnosticar a proporção entre notícias falsas e verdadeiras no conjunto de dados final. A quantificação detalhada e a representação visual dessa distribuição foram reportadas no Capítulo 5, demonstrando o equilíbrio alcançado após a limpeza dos dados e dispensando a necessidade de técnicas de reamostragem.

4.2.1 Visualização de Frequência e Definição dos Pontos de Mudança

Para visualizar a frequência dos termos mais relevantes no corpus e identificar os tópicos centrais que permeiam o conjunto de dados, utilizou-se a técnica de visualização por Nuvens de Palavras. Nessa representação, o tamanho de cada termo é diretamente proporcional à sua frequência no texto, permitindo uma identificação visual imediata dos assuntos dominantes após a remoção de *stop words*.

[1] Disponível em: https://huggingface.co/conjuntodedadoss/recogna-nlp/fakerecogna2-extrativa

É fundamental ressaltar uma distinção metodológica aplicada nesta fase: essa filtragem de *stop words* teve caráter exclusivamente visual, visando evitar a poluição gráfica nas figuras geradas e destacar os termos de maior carga semântica. Importante frisar que esse tratamento de remoção de palavras funcionais não foi aplicado aos dados de treinamento do modelo BERTimbau, uma vez que a arquitetura *Transformer* se beneficia da estrutura sintática completa das sentenças para o cálculo dos mecanismos de atenção.

A geração das visualizações foi estruturada em dois níveis de granularidade. Inicialmente, elaborou-se uma visualização global a partir da totalidade do *corpus*, permitindo uma visão panorâmica dos termos mais frequentes. Em um segundo nível, a análise foi segmentada temporalmente para investigar o fenômeno da deriva de conceito.

A definição das janelas temporais para essa segmentação fundamenta-se no estudo de Wanderley *et al.* (2025), que realizou uma análise em larga escala sobre o *FakeRecogna 2.0* e demonstrou que mudanças significativas em padrões tópicos e semânticos ocorrem em momentos específicos. Baseando-se na compatibilidade direta dos dados, este trabalho adota os três principais pontos de mudança identificados na literatura para guiar a geração das nuvens de palavras e, posteriormente, a análise de explicabilidade:

- **Drift 1 (Início da Pandemia):** Identificado em 14/03/2020 para notícias verdadeiras e 16/03/2020 para notícias falsas. Este ponto marca a inserção abrupta do vocabulário e dos temas relacionados à crise sanitária da COVID-19.
- **Drift 2 (Evolução do Cenário):** Identificado em 27/07/2020 para notícias verdadeiras e 17/09/2020 para notícias falsas, representando uma mudança nos tópicos de discussão pública conforme a pandemia evoluía.
- **Drift 3 (Ponto de Estabilização):** Identificado em 04/12/2020 para notícias verdadeiras e 06/12/2020 para notícias falsas.

Para cada uma dessas janelas temporais, o conteúdo foi segregado por classe, resultando em duas nuvens distintas por *drift*: uma exclusiva para notícias verdadeiras e outra para notícias falsas. Essa segmentação permite contrastar se os tópicos centrais da desinformação divergiram dos temas abordados pelo jornalismo profissional nesses momentos críticos.

4.3 Definição e Adaptação do Modelo

Para a execução da tarefa de classificação, adotou-se o modelo BERTimbau Base. A escolha por esta versão justifica-se pela sua eficiência no processamento do português brasileiro,

aliada à viabilidade computacional necessária para a execução das etapas metodológicas que serão detalhadas posteriormente.

4.3.1 Adaptação de Domínio via MLM

Visando especializar o modelo para o vocabulário e as nuances sintáticas características do contexto jornalístico brasileiro, realizou-se uma etapa preliminar de treinamento não supervisionado. O objetivo desta etapa é ajustar as representações internas do modelo para capturar o vocabulário, a sintaxe e os padrões semânticos predominantes no corpus *FakeRecogna 2.0*, antes de submetê-lo à tarefa final de classificação.

Nesta abordagem, o modelo recebe como entrada o texto completo das notícias disponíveis no conjunto de dados, sem a utilização dos rótulos de classe. Seguindo o protocolo padrão do BERT, aplicou-se uma probabilidade de mascaramento de 15%, onde *tokens* aleatórios do texto de entrada são ocultados para que o modelo aprenda a reconstruí-los com base no contexto bidirecional.

A configuração experimental desta etapa de pré-treinamento adotou um comprimento máximo de sequência de 512 tokens e uma probabilidade de mascaramento de 0,15. O treinamento foi realizado com um tamanho de lote de 16 e uma taxa de aprendizadode 5×10^{-5}. Embora configurado para um limite máximo de 100 épocas, implementou-se um mecanismo de parada antecipada (*early stopping*) com paciência de 5 épocas, monitorando a estagnação da função de perda no conjunto de validação. O critério de parada foi acionado na 32ª época, momento em que o modelo atingiu a convergência desejada, evitando o sobreajuste aos dados. Os pesos resultantes deste processo constituem o que este trabalho denomina de 'Modelo Pré-treinado'.

4.4 Pré-processamento

A etapa de pré-processamento foi estruturada para assegurar a integridade dos dados e a adequação aos requisitos de entrada do modelo para a tarefa de classificação supervisionada. Inicialmente, aplicou-se uma filtragem para remover amostras que apresentassem valores ausentes nas colunas essenciais para o experimento: "Noticia", "Label" e "Autor".

É importante destacar que este procedimento de limpeza foi aplicado exclusivamente à fase de ajuste fino. Para a etapa anterior de pré-treinamento via MLM, os dados foram utilizados em sua forma bruta, sem a exclusão de amostras baseada na ausência de dados. Essa

distinção deve-se à natureza não supervisionada do MLM, que não requer rótulos e se beneficia da exposição ao maior volume possível de texto para a calibração dos padrões linguísticos.

4.4.1 Estratégia de Validação e Tratamento de Autoria

A robustez da avaliação experimental em conjuntos de dados de notícias depende da independência estrita entre os dados de treino e teste. Em tarefas de classificação de notícias falsas, a literatura aponta que a sobreposição de fontes atua como um forte fator de confusão (Zhou *et al.*, 2021). Existe, portanto, um risco significativo de vazamento de dados por viés de autoria: modelos podem alcançar alta performance memorizando o mapeamento entre o site/autor e o rótulo, em vez de modelar a tarefa real de detecção de desinformação (Zhou *et al.*, 2021).

Embora a variável explícita "Autor" não seja fornecida como entrada para o modelo, abordagens baseadas em estilometria demonstram que classificadores são capazes de identificar a proveniência do texto através de padrões linguísticos intrínsecos (Schuster *et al.*, 2020). A estilometria fundamenta-se na extração de características estilísticas do texto escrito por meio de métodos estatísticos, analisando desde a frequência de palavras até estruturas sintáticas complexas para inferir a autoria ou detectar intenções enganosas. Caso notícias de uma mesma fonte fossem distribuídas aleatoriamente entre treino e validação, o modelo poderia aprender a correlacionar o estilo de escrita específico de uma agência com a classe alvo. Estudos demonstram que essa prática infla artificialmente as métricas de desempenho, com quedas de acurácia superiores a 10% quando a separação por fonte é aplicada corretamente (Zhou *et al.*, 2021).

Para mitigar esse risco e evitar que o modelo explore atalhos estilísticos, adotou-se o protocolo de validação cruzada *StratifiedGroupKFold* com 10 *folds*. Este método garante que todas as notícias produzidas por uma mesma fonte permaneçam agrupadas, sendo alocadas integralmente ou no conjunto de treinamento ou no de validação ou no de teste, mas nunca em mais de um simultaneamente. Isso obriga o modelo a generalizar o aprendizado para estilos de escrita nunca vistos anteriormente, simulando um cenário mais realista de aplicação (Zhou *et al.*, 2021).

A viabilidade técnica da estratégia de validação cruzada por grupos demandou a normalização prévia dos dados na coluna de autoria, uma vez que o agrupamento depende da correspondência exata dos identificadores textuais. Sem esse tratamento, simples variações de grafia levariam o algoritmo a interpretar a mesma fonte como autores distintos, comprometendo

o isolamento necessário entre treino e teste.

O processo de padronização atuou na unificação de entidades através da remoção de prefixos editoriais, como o termo “Por”, e da uniformização da caixa das letras. Essa medida garantiu, por exemplo, que variações de escrita como “Por G1” e “Por g1” fossem consolidadas sob um único identificador. Simultaneamente, procedeu-se à remoção de ruídos de extração onde a coluna de autor continha erroneamente apenas carimbos de data e hora, a exemplo de “15/12/2020 15h49”. Tais padrões foram detectados via expressões regulares e tratados para evitar a criação de grupos artificiais. Nos casos residuais em que o nome do autor resultou vazio após a limpeza, atribuiu-se um rótulo genérico de “Autor desconhecido”. Após a execução dessas rotinas de tratamento e padronização, o conjunto de dados final consolidado para os experimentos totalizou 49.988 instâncias.

4.5 Treinamento dos Modelos

Esta etapa compreende a fase de aprendizado supervisionado, na qual o modelo é submetido aos dados processados para aprender a distinguir entre notícias falsas e verdadeiras. A implementação foi realizada utilizando a biblioteca *Transformers* sobre o *framework* PyTorch, com o ambiente configurado para execução acelerada por GPU.

4.5.1 Congelamento de Camadas

Diferentemente do ajuste fino integral (*fine-tuning*), onde todos os parâmetros da rede são recalibrados, este trabalho adotou a estratégia de Extração de Características. Nesta abordagem, os pesos das 12 camadas internas do codificador BERTimbau Base (tanto na versão Referência quanto na Pré-treinada) foram mantidos congelados, de modo que o algoritmo de otimização atualizou exclusivamente os pesos da camada linear final de classificação.

Essa escolha metodológica fundamenta-se em dois pilares principais. Primeiramente, buscou-se o isolamento da qualidade das representações como etapa prévia à aplicação das técnicas de explicabilidade. Ao impedir a atualização do codificador, força-se o classificador a depender inteiramente da qualidade das representações vetoriais pré-existentes, permitindo verificar se a adaptação via MLM efetivamente produziu um espaço latente mais robusto. Simultaneamente, o congelamento proporciona eficiência computacional, viabilizando as múltiplas iterações exigidas pela validação cruzada.

4.5.2 Modelos Avaliados e Protocolo Experimental

Com o intuito de analisar o impacto da adaptação de domínio e fundamentar a escolha do classificador para a análise de explicabilidade, foram executados dois experimentos paralelos. Ambos os modelos foram submetidos rigorosamente ao mesmo protocolo de treinamento supervisionado descrito anteriormente, utilizando os mesmos 10 *folds* de validação cruzada.

Para a definição do fluxo de dados deste experimento, utilizou-se exclusivamente o conteúdo textual da coluna "Noticia" como variável de entrada para o codificador, enquanto a coluna "Label" serviu como variável alvo para o cálculo da função de perda e otimização do classificador linear.

As duas variações de arquitetura avaliadas foram:

- Modelo de Referência: Utiliza os pesos originais do BERTimbau Base, pré-treinado em um *corpus* genérico. O treinamento da camada de classificação iniciou-se diretamente sobre esses pesos, servindo como linha de base para verificar o desempenho sem o conhecimento específico do domínio.
- Modelo Pré-treinado (Com MLM): Utiliza os pesos refinados após a etapa de pré-treinamento via MLM no corpus *FakeRecogna 2.0*. O classificador linear foi treinado sobre essas representações especializadas, visando capturar o ganho proporcionado pelas nuances aprendidas na etapa prévia.

A validação da comparação foi assegurada pela manutenção de hiperparâmetros idênticos para ambos os experimentos. O classificador de cada *fold* foi treinado por um número fixo de 4 épocas, utilizando a mesma semente de aleatoriedade. O uso da mesma semente garante que a divisão dos subconjuntos de dados fossem idênticas em ambos os cenários. Esse rigor metodológico isola a qualidade das características geradas pelo pré-treinamento via MLM como a única variável significativa do experimento.

O desempenho final de cada abordagem foi determinado pela média e pelo desvio padrão das métricas de Acurácia, Precisão, Revocação e *F1-Score*, calculadas exclusivamente sobre os conjuntos de teste de cada *fold*.

4.6 Avaliação dos Resultados

Concluído o processo de treinamento e validação cruzada, a etapa final consiste na avaliação de desempenho dos modelos selecionados sobre os conjuntos de teste.

Para aferir a performance dos classificadores, foram reportadas as médias e os desvios padrão das métricas de Acurácia, Precisão, Revocação e F1-Score. A análise conjunta desses indicadores foi fundamental para validar se o pré-treinamento via MLM proporcionou ganhos reais na capacidade de classificação.

Adicionalmente, foi gerada a Matriz de Confusão dos *folds* que os modelos tiveram o melhor resultado. Essa ferramenta permitiu uma análise visual e qualitativa dos erros, detalhando as tendências de classificação do modelo.

Para a condução das análises subsequentes, optou-se pela seleção de uma única instância do classificador. Vale ressaltar que o protocolo de validação cruzada estratificada por grupos ($k = 10$) gerou, ao todo, dez versões distintas do modelo, uma para cada iteração de treino e validação. A estratificação por grupos utilizou a variável "Autor" estritamente como critério de isolamento, garantindo que todas as notícias de um mesmo autor fossem alocadas exclusivamente no conjunto de treino ou no de validação, evitando assim o vazamento de dados via estilometria. Dentre as dez instâncias resultantes desse processo, foi eleita aquela correspondente ao *fold* que apresentou os melhores resultados nas métricas de validação. Essa escolha estratégica visa assegurar que o estudo de caso com o LIME seja realizado sobre a versão mais robusta da arquitetura, refletindo seu potencial máximo de aprendizado sem enviesar a avaliação final.

4.6.1 Protocolo de Seleção e Aplicação do LIME

Para verificar a hipótese de que as técnicas de explicabilidade conseguem capturar e justificar as mudanças semânticas decorrentes da deriva de conceito, bem como permitir uma auditoria qualitativa do modelo, estabeleceu-se um protocolo experimental para cada um dos três pontos de mudança listados.

O procedimento inicia-se com a definição de uma janela temporal de observação de 7 dias ao redor do evento, compreendendo o dia do *drift*, acrescido de três dias anteriores e três dias posteriores, com o intuito de capturar o contexto imediato da mudança. Dentro dessa janela, realiza-se uma filtragem por confiança, selecionando as 10 amostras de cada classe (falsa e verdadeira) nas quais o modelo apresentou a maior probabilidade de predição, visando isolar

os padrões considerados mais característicos daquele período. Por fim, aplica-se o método LIME a essas amostras para gerar explicações locais, identificando quais *tokens* exerceram maior influência positiva ou negativa na decisão do classificador.

A análise subsequente das explicações geradas buscou validar se os termos destacados correspondem à emergência de novos vocabulários ou tópicos, diagnosticando a capacidade de adaptação temporal do modelo, além de permitir uma inspeção qualitativa sobre quais elementos linguísticos fundamentaram as decisões de alta confiança do classificador.

5 RESULTADOS

Neste capítulo, são apresentados os resultados obtidos a partir da aplicação da metodologia descrita no Capítulo 4.

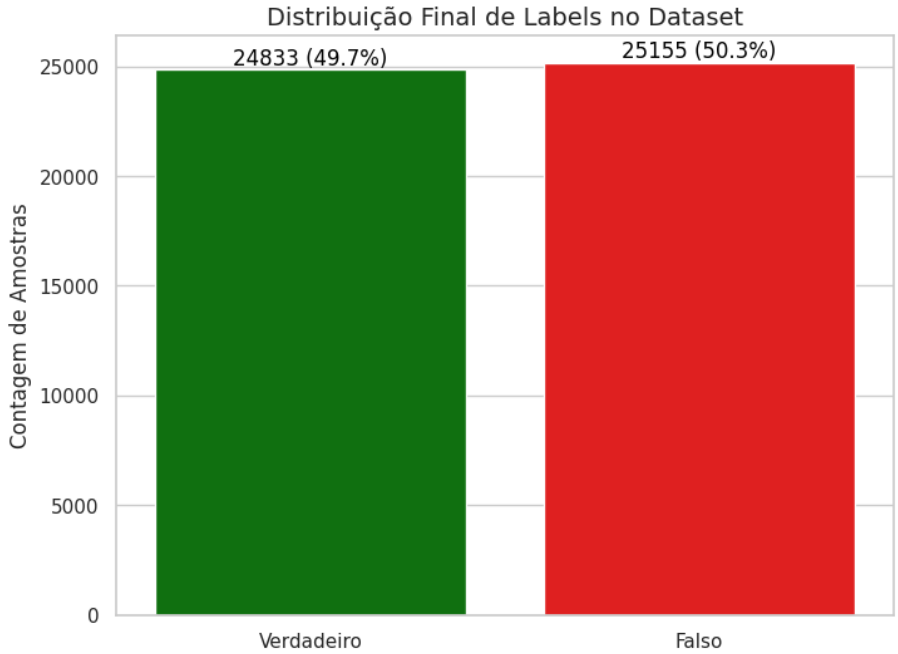
5.1 Análise Exploratória dos Dados

A aplicação combinada dos protocolos de limpeza dos dados e da estratégia de balanceamento baseada no perfil dos autores, conforme detalhado no Capítulo 4, resultou na consolidação do *corpus* final para o treinamento.

Partindo de um volume inicial de 52.773 entradas brutas, a remoção de registros inconsistentes reduziu primeiramente o conjunto para 51.279 amostras. Subsequentemente, a aplicação da subamostragem para equalização das fontes de autoria refinou o conjunto de dados para um total definitivo de 49.988 instâncias.

A distribuição final das classes neste conjunto processado é apresentada na Figura 6.

Figura 6 – Distribuição de classes no conjunto de dados final.



Fonte: Elaborado pelo autor.

Conforme evidenciado no gráfico, o conjunto de dados final exibe um equilíbrio estatístico, contendo 24.833 amostras rotuladas como verdadeiras (49,7%) e 25.155 como falsas (50,3%). Essa simetria confirma a eficácia da estratégia de subamostragem adotada, garantindo que o modelo seja treinado sobre uma distribuição balanceada e mitigando o risco de

enviesamento em direção a uma classe majoritária.

5.1.1 Nuvem de Palavras Global

A análise léxica consolidada de todo o *corpus* resultou na visualização apresentada na Figura 7. Esta representação exibe os termos de maior frequência global, oferecendo um panorama dos tópicos centrais abordados nas notícias coletadas, independentemente da classe ou do período temporal.

Figura 7 – Nuvem de palavras global do conjunto de dados

Fonte: Elaborado pelo autor.

A análise visual revela uma forte predominância de termos entrelaçados aos contextos sanitário e político brasileiro. Vocábulos como “covid”, “pandemia”, “saúde”, “vacina” e “vírus” aparecem com grande destaque, corroborando a influência massiva da crise sanitária da COVID-19 no fluxo de informações contido no *FakeRecogna 2.0*.

Simultaneamente, a presença marcante de termos como “brasil”, “governo”, “bolsonaro”, “jair” e “presidente” evidencia a politização dos temas abordados. Essa interseção visual entre vocabulário de saúde pública e figuras políticas sugere que uma parcela significativa das notícias, tanto verdadeiras quanto falsas, orbita as discussões e polarizações ocorridas durante a gestão da pandemia no Brasil.

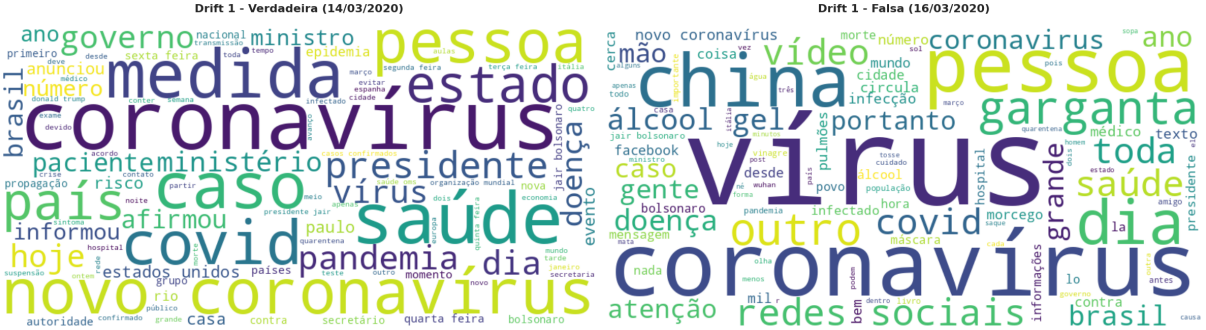
Tal predominância temática é estatisticamente justificada pela distribuição temporal do conjunto de dados. Conforme a análise da deriva de conceito realizada por Wanderley

et al. (2025) sobre este mesmo conjunto de dados, observa-se uma concentração massiva de publicações nos anos de 2020 e 2021. O estudo aponta que aproximadamente 98% de todas as notícias falsas presentes no conjunto de dados datam desse período específico, o que explica a hegemonia absoluta dos termos relacionados ao auge da crise sanitária e aos seus desdobramentos políticos na nuvem de palavras global.

5.1.2 Análise Comparativa: Drift 1 - Início da Pandemia

A primeira janela de mudança de conceito situa-se em meados de março de 2020 (dias 14 e 16), coincidindo com a declaração oficial da pandemia pela OMS. Para visualizar as divergências semânticas neste momento crítico, a Figura 8 exibe as nuvens de palavras das notícias verdadeiras e falsas lado a lado.

Figura 8 – Nuvem de palavras no Drift 1



Fonte: Elaborado pelo autor.

À esquerda, as palavras das notícias verdadeiras refletem a reação institucional e administrativa. Termos como “medida”, “anunciou”, “ministério”, “governo” e “decreto” indicam um jornalismo focado em comunicar as primeiras ações de contenção do Estado e o monitoramento estatístico.

Em contrapartida, a nuvem das notícias falsas, à direita, desloca o foco para soluções milagrosas e teorias de origem. O destaque para palavras como “china” e “morcego” evidencia a força imediata das narrativas conspiratórias e xenofóbicas sobre a criação do vírus. Simultaneamente, a presença de termos anatômicos e domésticos como “garganta”, “água”, “vinagre”, “álcool” e “matar” remete à proliferação de receitas falsas que prometiam a cura ou prevenção caseira da doença, como os boatos de que fazer gargarejos com vinagre ou beber água morna eliminaria o vírus, disseminadas amplamente em redes sociais, notadas pelos termos “facebook” e “mensagem”.

5.1.3 Análise Comparativa: Drift 2 - Evolução do Cenário

O segundo ponto de mudança de conceito destaca-se pelo descompasso temporal e temático entre as classes. Conforme identificado na análise estatística de Wanderley *et al.* (2025), este *drift* ocorre em momentos distintos: 27 de julho de 2020 para as notícias verdadeiras e 17 de setembro de 2020 para as notícias falsas. A Figura 9 ilustra as nuvens de palavras correspondentes a essas janelas.

Figura 9 – Nuvem de palavras no Drift 2

Fonte: Elaborado pelo autor.

Na nuvem das notícias verdadeiras, observa-se uma estabilidade semântica. Mesmo quatro meses após o início da pandemia, o vocabulário permanece ancorado na crise sanitária, com predominância absoluta de termos como “covid”, “pandemia”, “saúde” e “mortes”. A presença de palavras como “mil” e “número” sugere o estágio de contagem de vítimas e a consolidação das estatísticas da doença. O jornalismo, portanto, manteve sua pauta prioritária fixa na emergência global.

Em contraste, a nuvem das notícias falsas revela a natureza oportunista e volátil da desinformação. O termo “covid” perde relevância e cede espaço para um novo tópico emergente: os desastres ambientais. Palavras como “fogo”, “amazônia”, “incêndio”, “pantanal” e “queimada” dominam o léxico, refletindo o período de queimadas recordes no Brasil em meados de 2020.

A análise visual destaca a enorme dimensão do termo “facebook”, indicando o canal primário dessa onda de desinformação. Além disso, chama atenção a proeminência do verbo “circula”. Este termo atua como um validador social típico de boatos (ex: “circula nas redes”, “vídeo que circula”), criando uma falsa sensação de urgência e popularidade. Juntamente com a politização das tragédias, evidenciada por “lula”, “dilma” e “governo”, essa transição demonstra que os vetores de notícias falsas migram rapidamente de assunto, explorando a polêmica do momento para manter o engajamento.

5.1.4 Análise Comparativa: Drift 3 - Ponto de Estabilização

O terceiro e último ponto de análise temporal situa-se no início de dezembro de 2020. Segundo Wanderley *et al.* (2025), este momento representa uma "estabilização" da deriva de conceito, onde os padrões de desinformação se consolidam. A Figura 10 exibe as nuvens de palavras para as notícias verdadeiras (04/12) e falsas (06/12).

Figura 10 – Nuvem de palavras no Drift 3

Fonte: Elaborado pelo autor.

Neste estágio, observa-se uma convergência temática: a palavra "vacina" torna-se o termo central em ambas as classes. No entanto, o contexto semântico permanece antagônico.

Nas notícias verdadeiras, o léxico reflete o início das campanhas globais de imunização. Termos como "Pfizer", "Reino Unido", "imunizante", "plano" e "anvisa" denotam um discurso focado na logística, na aprovação regulatória e na ciência. A presença das palavras "estado" e "governo" reafirma o caráter institucional da cobertura.

Visualmente, o que mais impacta nesta nuvem de notícias falsas é a desproporção dos vetores de disseminação. As palavras "Vídeo" e "WhatsApp" assumem grandes dimensões, superiores até mesmo a muitos termos de conteúdo. Isso sugere que, neste ponto, o meio de propagação tornou-se parte intrínseca da mensagem.

Novamente, o verbo "circula" aparece com grande destaque, consolidando-se como um marcador linguístico perene. Diferente da notícia jornalística que relata o fato ("ocorreu", "anunciou"), a desinformação relata o seu próprio movimento ("circula", "compartilhem"), utilizando essa autorreferência para persuadir o leitor.

5.2 Análise de Desempenho dos Classificadores

A avaliação quantitativa dos modelos foi conduzida através da validação cruzada estratificada por grupos em 10 rodadas. O objetivo primário foi comparar a capacidade de

generalização do modelo de referência (BERTimbau sem pré-treinamento com camadas congeladas) contra o modelo proposto (BERTimbau com pré-treinamento via MLM e com camadas congeladas).

Os resultados obtidos evidenciaram a superioridade de desempenho da abordagem proposta frente ao modelo de referência. A Tabela 1 exibe um resumo comparativo das médias e desvios padrão das métricas de avaliação ao longo dos 10 *folds* de validação, demonstrando um incremento consistente nas métricas obtidas pelo modelo submetido ao pré-treinamento.

Tabela 1 – Comparativo detalhado das métricas de validação entre o Modelo de Referência e o Modelo Pré-treinado.

	Acurácia		**Precisão**		**Revocação**		***F1-Score***	
fold	**Referência**	**Pré-treinado**	**Referência**	**Pré-treinado**	**Referência**	**Pré-treinado**	**Referência**	**Pré-treinado**
1	0,8472	**0,9623**	0,9797	**0,9970**	0,8230	**0,9551**	0,8945	**0,9756**
2	0,8519	**0,9221**	0,9790	**0,9946**	0,8297	**0,9060**	0,8982	**0,9482**
3	0,8401	**0,9629**	0,9805	**0,9962**	0,8131	**0,9565**	0,8890	**0,9760**
4	0,8295	**0,9588**	0,9807	**0,9960**	0,7992	**0,9514**	0,8807	**0,9732**
5	0,8338	**0,9563**	0,9796	**0,9962**	0,8056	**0,9482**	0,8841	**0,9716**
6	0,8518	**0,9603**	0,9790	**0,9953**	0,8296	**0,9540**	0,8981	**0,9742**
7	0,8330	**0,9457**	0,9808	**0,9969**	0,8036	**0,9339**	0,8834	**0,9644**
8	0,8462	**0,9514**	0,9791	**0,9964**	0,8221	**0,9416**	0,8938	**0,9683**
9	0,8407	**0,9604**	0,9801	**0,9965**	0,8141	**0,9530**	0,8894	**0,9743**
10	0,8841	**0,9717**	0,9411	**0,9849**	0,9097	**0,9791**	0,9251	**0,9820**
Média	0,8458	**0,9552**	0,9760	**0,9950**	0,8250	**0,9479**	0,8936	**0,9708**
Desvio	(± 0,0147)	(± 0,0128)	(± 0,0116)	(± 0,0034)	(± 0,0300)	(± 0,0178)	(± 0,0120)	(± 0,0087)

Fonte: Elaborado pelo autor.

A análise dos dados evidencia que, embora o modelo de referência apresente um desempenho robusto, ele ainda demonstra um certo desequilíbrio entre as métricas: sua alta Precisão média (0,9760) contrasta com uma Revocação de 0,8250. Em termos práticos, isso indica que o classificador base, sem adaptação, tende a ser mais conservador, falhando em detectar cerca de 17,5% das notícias falsas reais presentes no conjunto de teste (Falsos Negativos).

O modelo pré-treinado superou a referência em todas as métricas consolidadas. O ganho mais expressivo ocorreu na Revocação, que saltou de 0,8250 para 0,9479, demonstrando que a especialização no domínio permitiu ao classificador identificar a vasta maioria das notícias falsas anteriormente ignoradas. Importante notar que este ganho de sensibilidade não gerou um *trade-off* com a Precisão, que atingiu o patamar de 0,9950. O modelo tornou-se, portanto, mais sensível e específico simultaneamente, elevando o *F1-Score* de 0,8936 para 0,9708. Além disso, o desvio padrão menor nas métricas do modelo pré-treinado indica uma maior robustez e estabilidade na generalização entre diferentes grupos de autores.

A etapa anterior de pré-treinamento via MLM utilizou o *corpus* completo. Embora

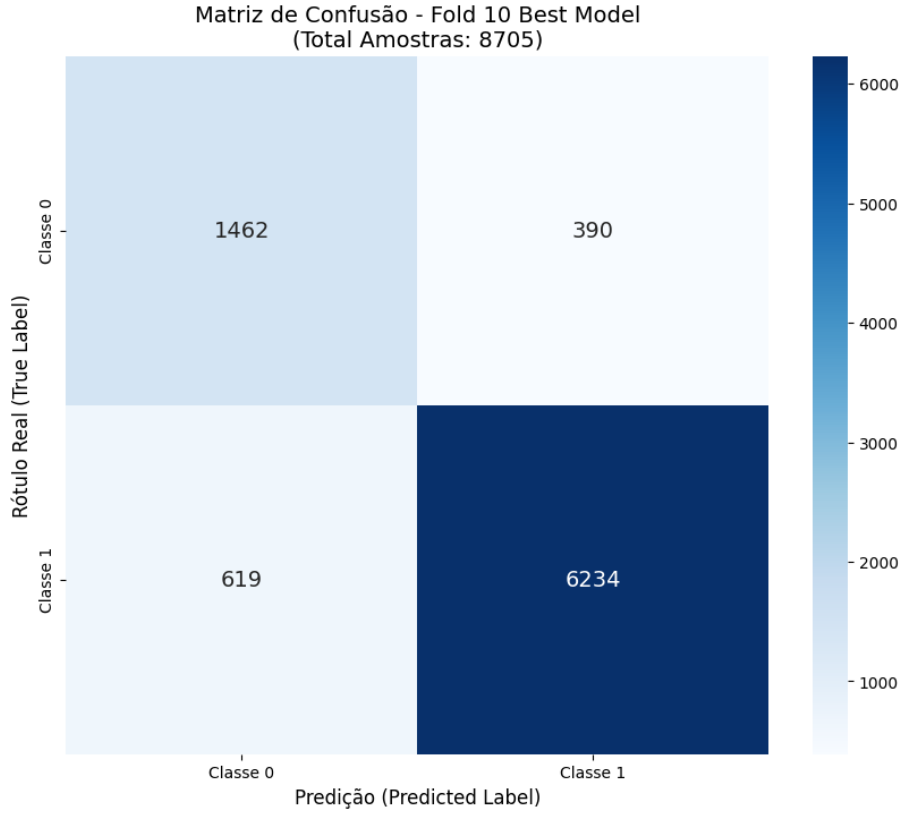
isso configure tecnicamente uma exposição não supervisionada aos textos de validação e teste, é fundamental pontuar que os rótulos de classificação permaneceram estritamente isolados. Dessa forma, o modelo não teve acesso à "resposta" final durante o pré-treinamento, apenas adaptou-se à estrutura estatística e ao vocabulário do domínio, o que mitiga o impacto dessa limitação na validade dos resultados comparativos.

5.2.1 Comparativo das Matrizes de Confusão

Para compreender qualitativamente o impacto do pré-treinamento na decisão do classificador, comparam-se as matrizes de confusão do Modelo de Referência e do Modelo Pré-treinado. Ambas correspondem ao desempenho no *Fold* 10, cenário de melhor generalização observado para ambos os casos.

A Figura 11 exibe os resultados do Modelo de Referência. A análise visual confirma a limitação de sensibilidade apontada pelas métricas: o classificador gerou 619 Falsos Negativos. Isso significa que, sem o pré-treinamento via MLM, o modelo falhou em identificar uma quantidade significativa de notícias falsas, rotulando-as incorretamente como verdadeiras. Além disso, observa-se uma taxa considerável de 390 Falsos Positivos.

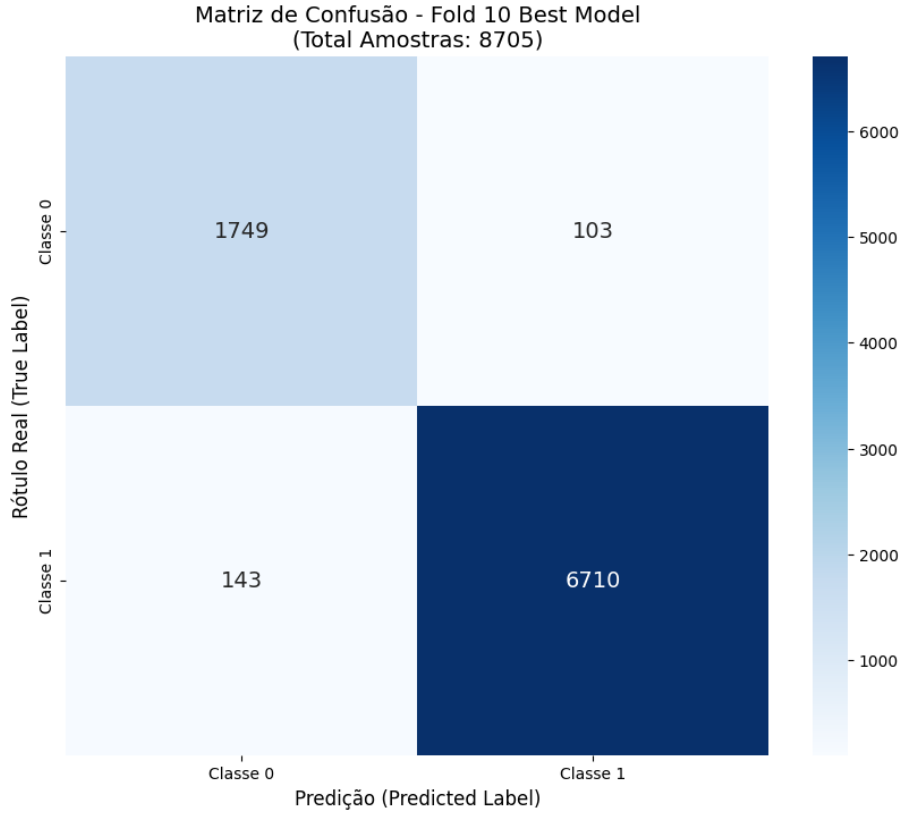
Figura 11 – Matriz de Confusão - Modelo de Referência (*Fold* 10)



Fonte: Elaborado pelo autor.

Em contrapartida, a Figura 12 exibe a matriz do Modelo Pré-treinado, evidenciando a evolução do classificador.

Figura 12 – Matriz de Confusão - Modelo Pré-treinado (*Fold* 10)



Fonte: Elaborado pelo autor.

A comparação direta revela o ganho de robustez. O número de Falsos Negativos caiu drasticamente de 619 para 143, confirmando que o modelo pré-treinado tornou-se muito mais eficaz em detectar as notícias falsas. Simultaneamente, a precisão também aumentou, visto que os Falsos Positivos foram reduzidos de 390 para apenas 103.

O resultado final é um classificador que acerta mais em ambas as classes: os VP subiram de 6.234 para 6.710, enquanto os Verdadeiros Negativos aumentaram de 1.462 para 1.749. Esses números corroboram a hipótese de que a adaptação de domínio via MLM refinou a representação vetorial do BERT, permitindo a distinção de nuances linguísticas sutis que passavam despercebidas pelo modelo de referência.

5.3 Análise de Interpretabilidade

O objetivo desta etapa foi verificar se o ganho de desempenho obtido pelo modelo pré-treinado reflete uma "atenção"mais coerente aos termos chave do domínio, em contraste com

possíveis vieses ou padrões que os modelos possam ter aprendido.

Esta investigação visou atender diretamente ao objetivo específico de aplicar a técnica LIME para gerar explicações locais das decisões dos classificadores. Ao confrontar as justificativas do modelo de referência com as do modelo pré-treinado, buscou-se validar se o processo de pré-treinamento promoveu alterações nos padrões linguísticos e no vocabulário considerados determinantes para a classificação, permitindo compreender não apenas se o modelo melhorou, mas como seus critérios de decisão evoluíram.

A estratégia de seleção das amostras para esta análise baseou-se no nível de confiança da predição. Para cada um dos três momentos de mudança de conceito identificados anteriormente, selecionou-se a instância que cada modelo classificou com a maior probabilidade para a classe alvo. Essa abordagem permite visualizar o "arquétipo"do que cada classificador considera, com máxima certeza, uma notícia verdadeira ou falsa naquele contexto temporal específico.

5.3.1 Análise do Drift 1

A primeira janela de análise, denominada *Drift* 1, situa-se em março de 2020, marco temporal que delimita o início da crise sanitária global e a consequente introdução abrupta do vocabulário pandêmico no debate público.

5.3.1.1 Notícias Verdadeiras

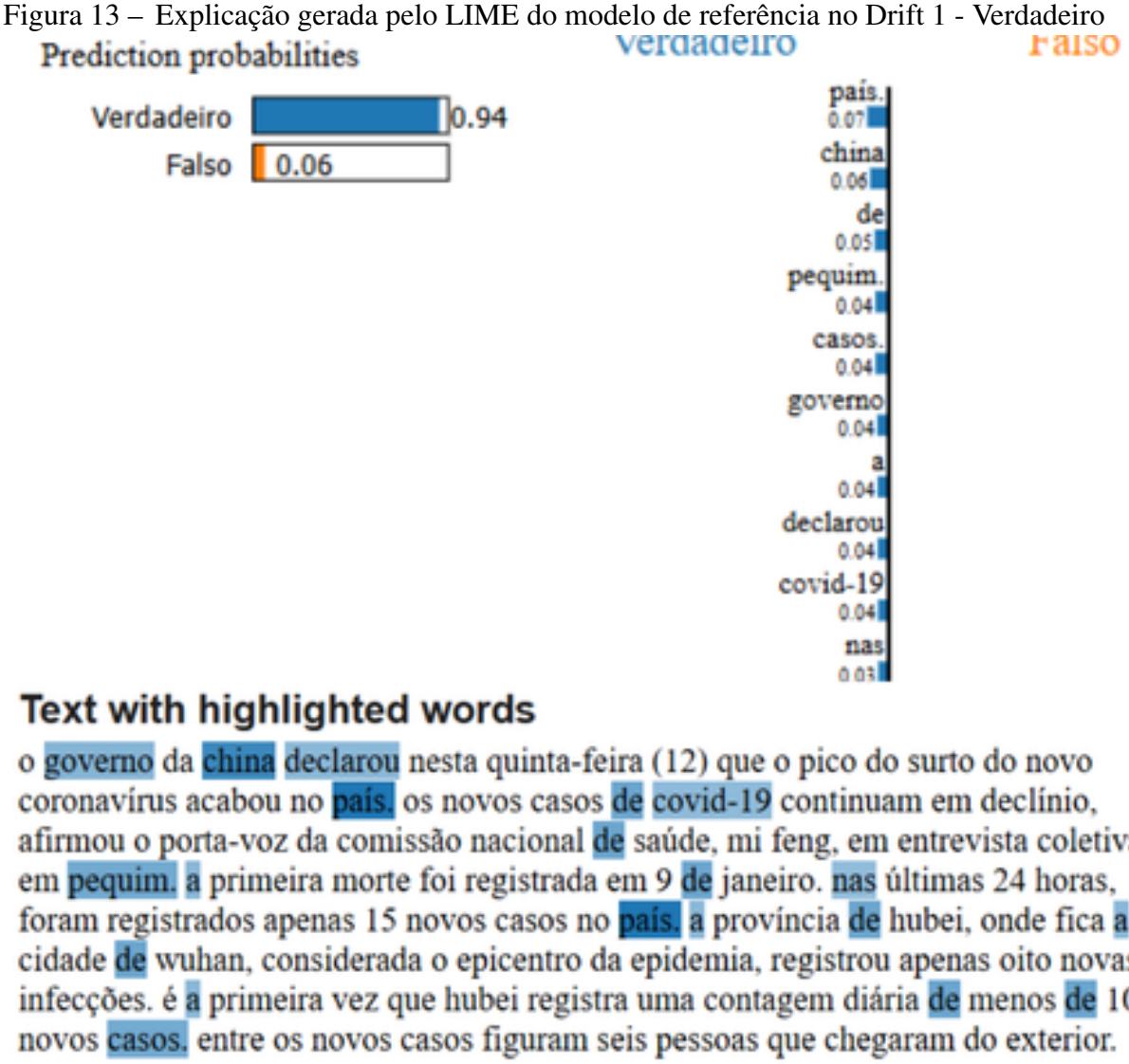
A Figura 13 exibe a explicação do Modelo de Referência para a amostra verdadeira classificada corretamente com uma confiança de 94,21%. O texto reporta declarações oficiais do governo chinês sobre o fim do pico do surto no país.

A análise detalhada dos *tokens* destacados revela os critérios utilizados pelo classificador base:

- “país”, “china”, “pequim” e “governo”: Estes termos receberam pontuações positivas. Isso pode indicar que o modelo de referência associa a veracidade à presença de entidades geográficas específicas e figuras de autoridade institucional.

- “de” e “a”: Observa-se que palavras funcionais comuns na língua portuguesa (*stopwords*) foram destacadas com pesos positivos. Embora isoladamente não carreguem conteúdo semântico específico, sua presença valorizada pelo modelo sugere que a estrutura sintática

Figura 13 – Explicação gerada pelo LIME do modelo de referência no Drift 1 - Verdadeiro



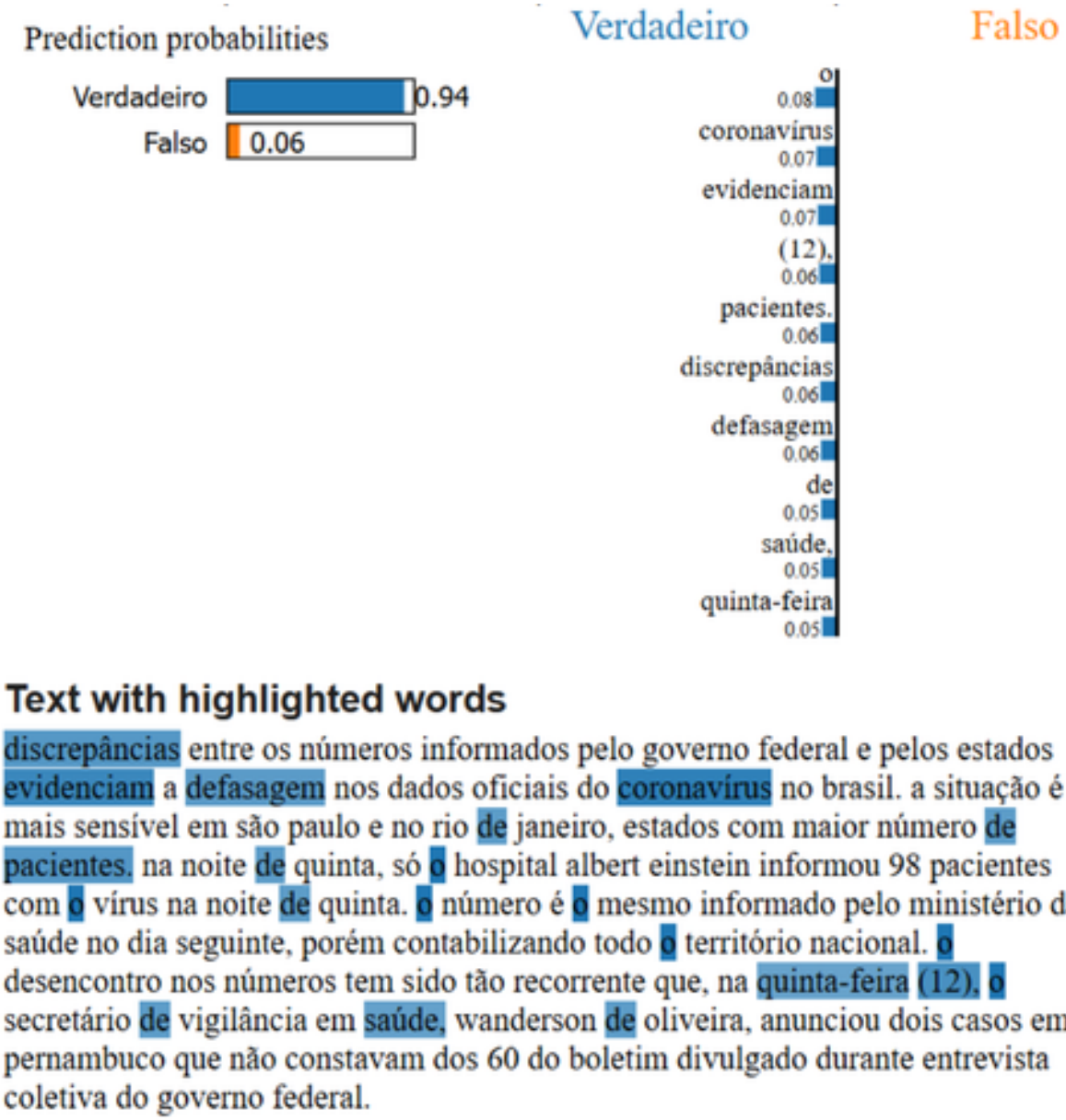
Fonte: Elaborado pelo autor.

correta atua como um preditor de veracidade. Isso indica que notícias reais tendem a ser gramaticalmente mais completas e bem escritas, enquanto a ausência desses conectivos pode ser uma característica estrutural de textos enganosos ou de baixa qualidade.

- “declarou” e “casos”: O verbo de elocução, típico do jornalismo declaratório, e o substantivo central da pandemia foram corretamente identificados como traços de veracidade, indicando que o modelo captou parcialmente o estilo de reportagem.

Já para o Modelo Pré-treinado, a Figura 14 exibe a interpretação para a amostra com uma confiança de 93,54%. O texto aborda um tema analítico sobre inconsistências estatísticas entre dados estaduais e federais.

Figura 14 – Explicação gerada pelo LIME do modelo pré-treinado no Drift 1 - Verdadeiro



Fonte: Elaborado pelo autor.

A análise das *features* revela uma mudança qualitativa na atenção do classificador em comparação ao modelo de referência:

- “discrepâncias”, “evidenciam” e “defasagem”: Diferente de textos falsos que apelam para linguagem emotiva ou simplista, estes termos atuam como assinaturas de um discurso analítico e ponderado, características intrínsecas ao jornalismo profissional que o modelo pode ter aprendido a valorizar.

- Referência ao contexto sanitário: as palavras “coronavírus” e “pacientes”, que foram desta-

cadas com pontuações positivas, constituem o vocabulário técnico padrão para a cobertura da pandemia. A presença desses termos denota objetividade e foco na descrição do evento e dos sujeitos afetados, características típicas de reportagens jornalísticas informativas, distanciando o texto de narrativas sensacionalistas ou politizadas.

- Palavras funcionais ("o"e "de"): são artigos e preposições extremamente comuns na língua portuguesa, frequentemente tratadas como *stopwords* devido à baixa carga semântica, mas que neste contexto receberam pontuação positiva. A relevância atribuída a esses termos pode sugerir que a estrutura gramatical normativa e o uso frequente de conectivos são características predominantes em notícias verdadeiras. Isso pode levar a considerar que o modelo associa a integridade sintática e a coesão textual a uma maior probabilidade de veracidade, diferenciando-a de textos virais que, por vezes, utilizam linguagem mais telegráfica ou informal.

5.3.1.2 Notícias Falsas

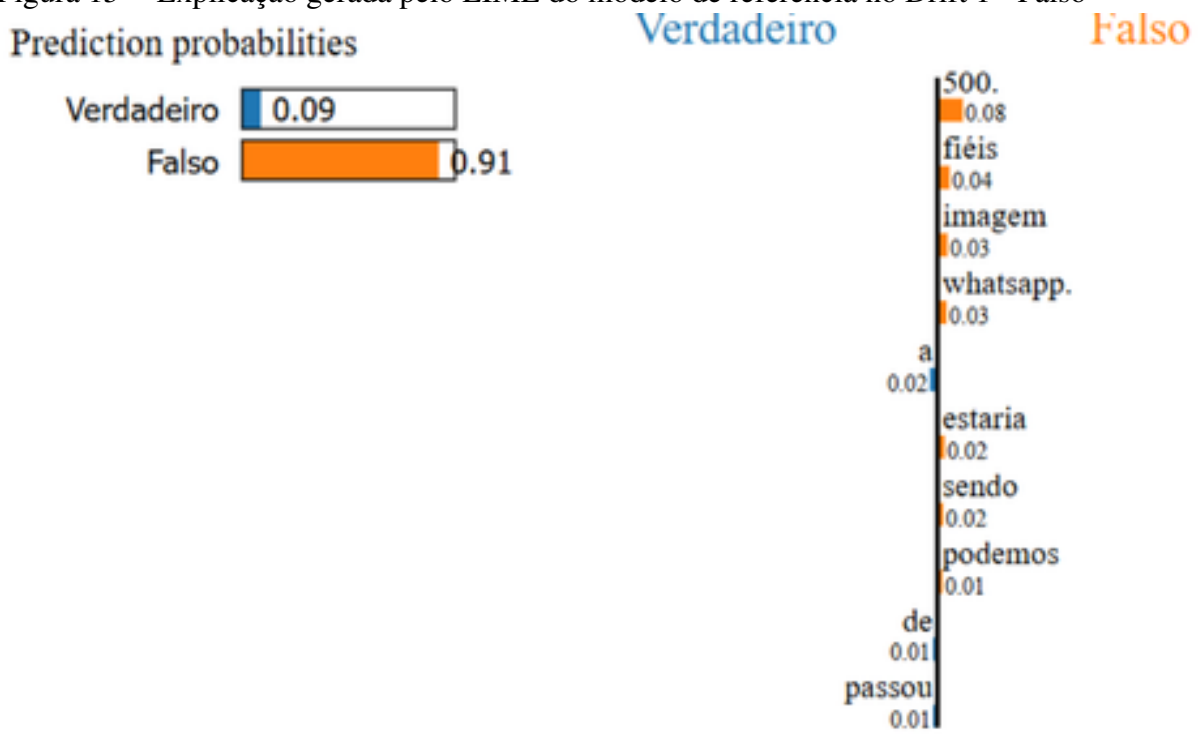
A Figura 15 exibe a explicação do Modelo de Referência para uma notícia falsa típica deste período, envolvendo boatos sobre curas milagrosas ou golpes financeiros.

A análise dos termos que mais contribuíram para a classificação de falsidade revela que o modelo capturou padrões associados a boatos virais e sensacionalismo:

- Público-alvo e viés numérico (“fiéis”, “500”): A palavra “fiéis” recebeu pontuação negativa relevante, sugerindo que o modelo pode ter identificado narrativas direcionadas a grupos religiosos como um padrão recorrente em desinformação. Já o destaque para o valor “500” pode sinalizar um viés do classificador em relação a cifras específicas, indicando uma correlação estatística aprendida pelo modelo com esse numeral, mais do que uma compreensão semântica do contexto financeiro.

- Referência ao meio de propagação (“whatsapp”, “imagem”): A presença explícita do nome do aplicativo de mensagens e a referência ao formato da mídia atuaram como preditores de falsidade. Isso indica que o classificador pode ter aprendido a associar a menção à própria viralização como uma característica de boatos, distinguindo-se de notícias profissionais que raramente citam o meio de transmissão como foco da notícia.

Figura 15 – Explicação gerada pelo LIME do modelo de referência no Drift 1 - Falso



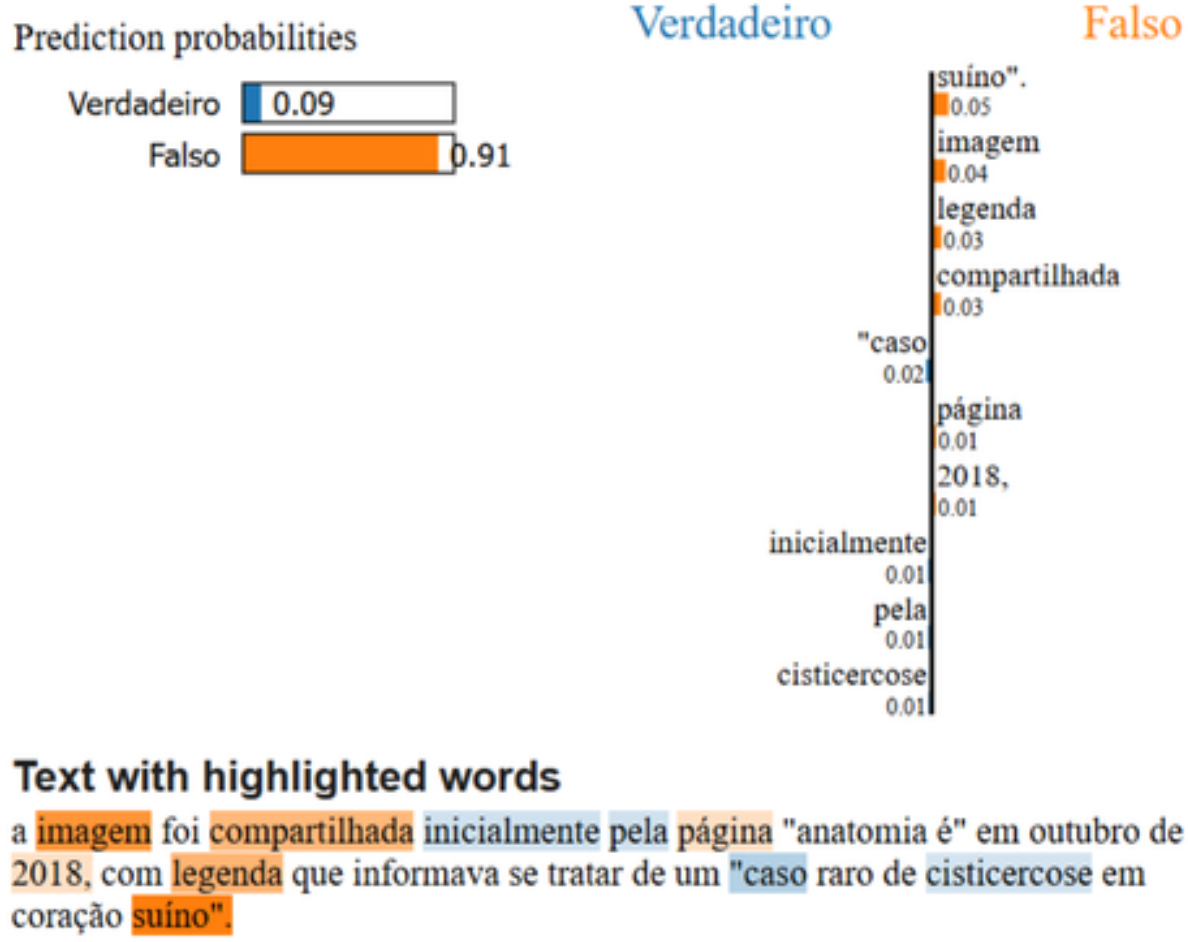
Fonte: Elaborado pelo autor.

A Figura 16 exibe a interpretação do Modelo Pré-treinado para a amostra falsa. O texto refere-se a um boato antigo sobre saúde (doença em carne suína) que voltou a circular.

- Vocabulário de viralização ("imagem", "compartilhada", "legenda"): Estes termos receberam pontuações negativas expressivas. Referênciam ao compartilhamento de notícias em redes sociais, isso pode se dar ao fato de que as notícias falsas são muito compartilhadas por meio dessas redes.

- "suíno": A palavra "suíno" recebeu o maior peso para a classificação falsa. Neste contexto, o termo define o objeto central de uma narrativa de pânico sanitário. Desinformações sobre saúde frequentemente utilizam alimentos de consumo popular (como a carne suína) como supostos vetores de doenças graves para gerar repulsa imediata e medo na população, uma

Figura 16 – Explicação gerada pelo LIME do modelo pré-treinado no Drift 1 - Falso



Fonte: Elaborado pelo autor.

tática comum para impulsionar a viralização de conteúdos sensacionalistas.

- "cisticercose": A atribuição de peso positivo a este termo médico segue a tendência observada em notícias reais, onde o vocabulário técnico de saúde é predominante. Neste caso, o termo técnico atua como um elemento de ancoragem de fatos que sugere legitimidade, embora o contexto geral da mensagem seja de boato.

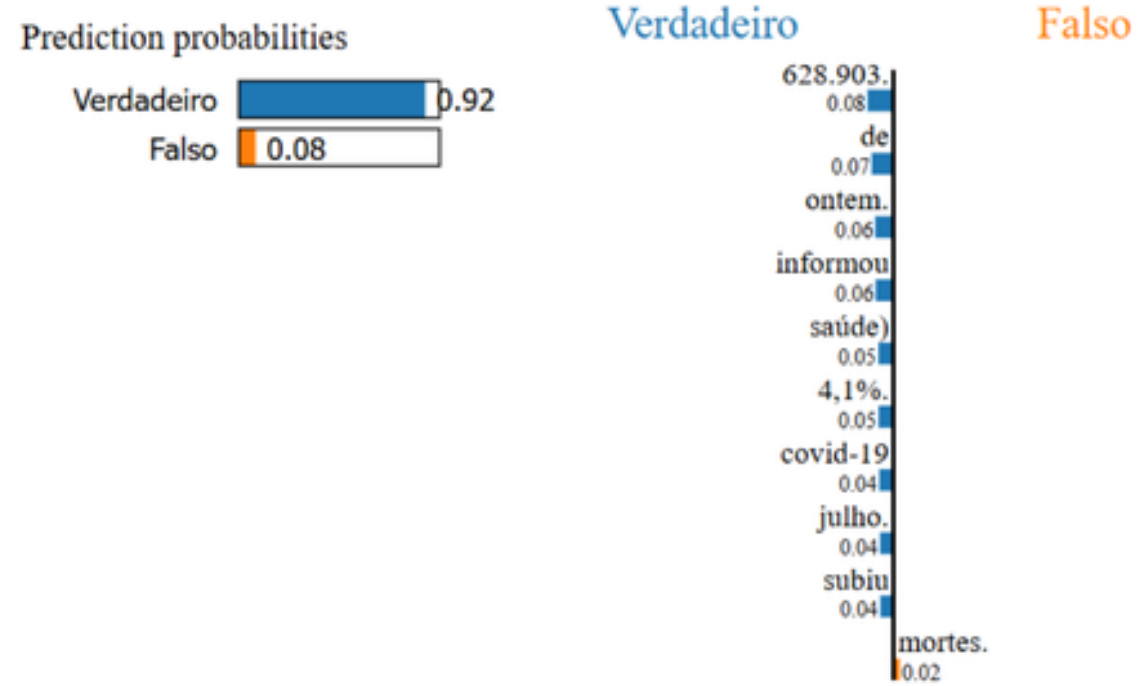
5.3.2 Análise do Drift 2

A segunda janela de análise, denominada *Drift* 2, refere-se ao estágio de evolução do cenário pandêmico, situado em meados de 2020.

5.3.2.1 Notícias Verdadeiras

A Figura 17 exibe a explicação do Modelo de Referência para a amostra verdadeira. O texto é um relatório estatístico baseado em dados da OMS sobre o número de mortos e casos confirmados, típico do estágio intermediário da pandemia.

Figura 17 – Explicação gerada pelo LIME do modelo de referência no Drift 2 - Verdadeiro



Text with highlighted words

o relatório mais recente da oms (organização mundial da saúde) informou que o número de mortos em decorrência do novo coronavírus ao redor do mundo subiu para 628.903. os dados foram compilados com informações recebidas pela organização até as 5h (de brasília) do dia 24 de julho. houve um aumento de 9.753 mortes em relação às informações disponibilizadas ontem. a quantidade de casos de covid-19 confirmados oficialmente aumentou para 15.296.926 casos. consideradas as informações disponíveis, a taxa global de mortalidade dos casos confirmados de coronavírus é de 4,1%. neste mesmo intervalo de tempo, peru (3.876), brasil (1.284) e estados unidos (1.074) foram os que registraram o maior número de novas mortes.

Fonte: Elaborado pelo autor.

- Precisão numérica e estatística (“628.903”, “4,1%”): Esses números tiveram peso positivo. A presença desses dados quantitativos específicos atua como um marcador de credibilidade, diferenciando o texto de boatos que geralmente utilizam números arredondados, estimativas vagas ou exageros sem fonte definida.

- Verbos de reportagem e marcação temporal (“informou”, “subiu”, “ontem”, “julho”): O verbo “informou” destaca-se por conferir credibilidade imediata ao texto, pois pressupõe a existência de uma fonte oficial ou organização responsável pela origem do dado, prática essencial no jornalismo profissional. Somado aos marcadores temporais que deram peso positivo para a predição.

- “de”: Novamente, a preposição recebeu um peso elevado.

Em contraste, a Figura 18 exibe a interpretação do Modelo Pré-treinado para a sua amostra de maior confiança neste mesmo período. Diferente do relatório estatístico analisado anteriormente, este texto aborda o impacto social da pandemia através de um relato pessoal (história de vida), um formato comum em reportagens de profundidade.

- Vocabulário de impacto social e rotina (“demissão”, “trajeto”, “agente”): Estes termos referem-se a aspectos da vida cotidiana e das relações trabalhistas. A sua relevância positiva sugere que a descrição detalhada de cenários reais pode atuar como peso positivo. Esse tipo de detalhamento é característico de reportagens que documentam os efeitos práticos da crise na vida dos cidadãos, distanciando-se da abstração ou do alarmismo vago comuns em boatos.

- Especificidade e identificação da fonte (“Tomaz”, “25”): A atribuição de relevância ao sobrenome da entrevistada e à sua idade reflete um padrão jornalístico de identificação precisa das fontes. Enquanto boatos costumam utilizar sujeitos indefinidos ("uma mulher disse", "um médico afirmou"), a notícia verdadeira ancora o relato em uma pessoa real com identidade definida, característica captada pelo modelo.

- “coronavírus”: Referências a esse termo têm peso positivo.

5.3.2.2 Notícias Falsas

A Figura 19 exibe a interpretação do Modelo de Referência para a amostra falsa. O texto descreve uma montagem ofensiva envolvendo uma figura política e um artista, ilustrando o

Figura 18 – Explicação gerada pelo LIME do modelo pré-treinado no Drift 2 - Verdadeiro

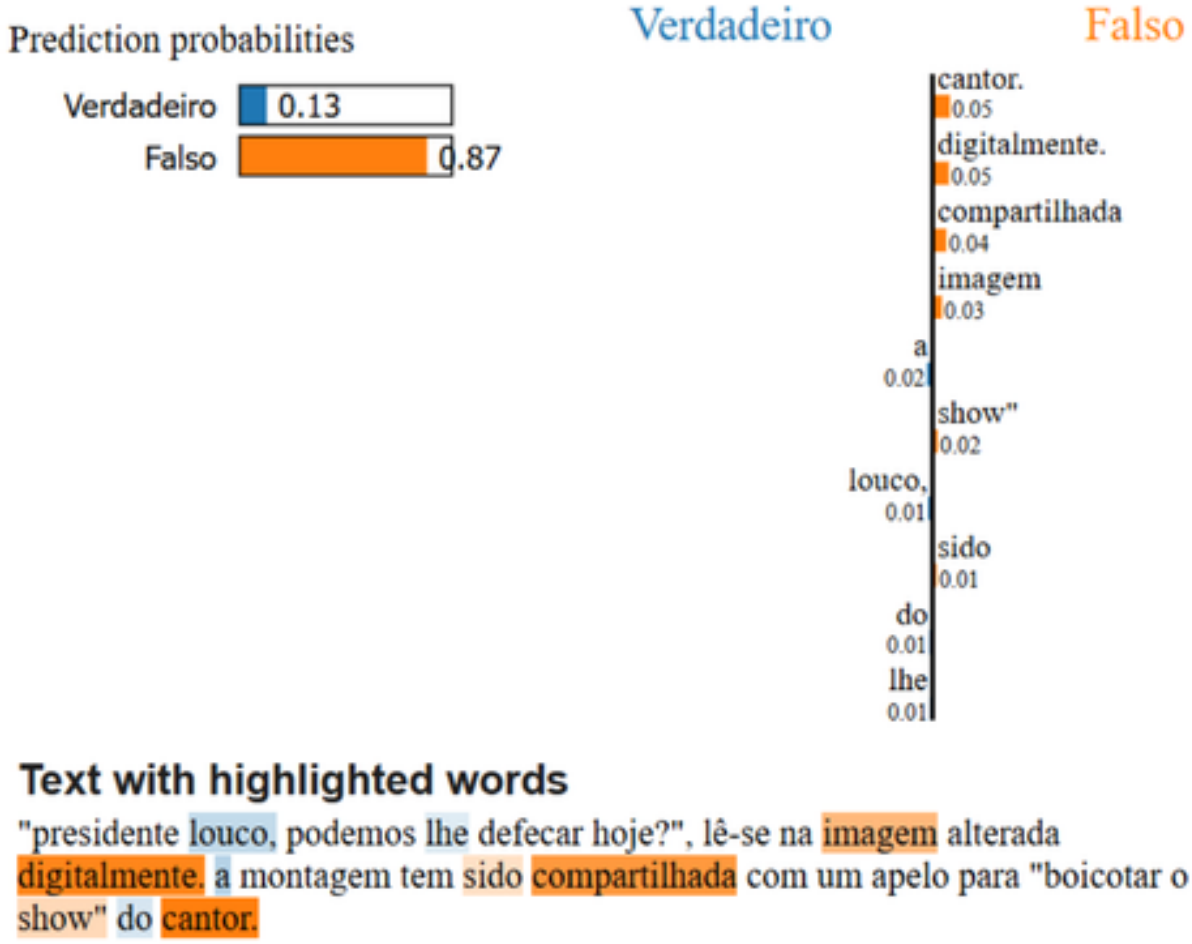
Fonte: Elaborado pelo autor.

uso de pautas de costumes e ataques culturais, comuns em estágios mais avançados da polarização online.

- Vocabulário de manipulação e viralização (“digitalmente”, “imagem”, “compartilhada”): Estes termos descrevem explicitamente a natureza artificial do conteúdo. A presença de palavras que denotam edição técnica ou a circulação de mídias visuais sinaliza que o texto trata de um conteúdo fabricado ou de um meme, afastando-se da narrativa de eventos físicos típica do jornalismo.

- Contexto de entretenimento (“cantor”, “show”): A relevância desses substantivos indica um deslocamento temático da desinformação para a esfera cultural. Notícias falsas frequen-

Figura 19 – LIME - Explicação gerada pelo LIME do modelo de referência no Drift 2 - Falso



Fonte: Elaborado pelo autor.

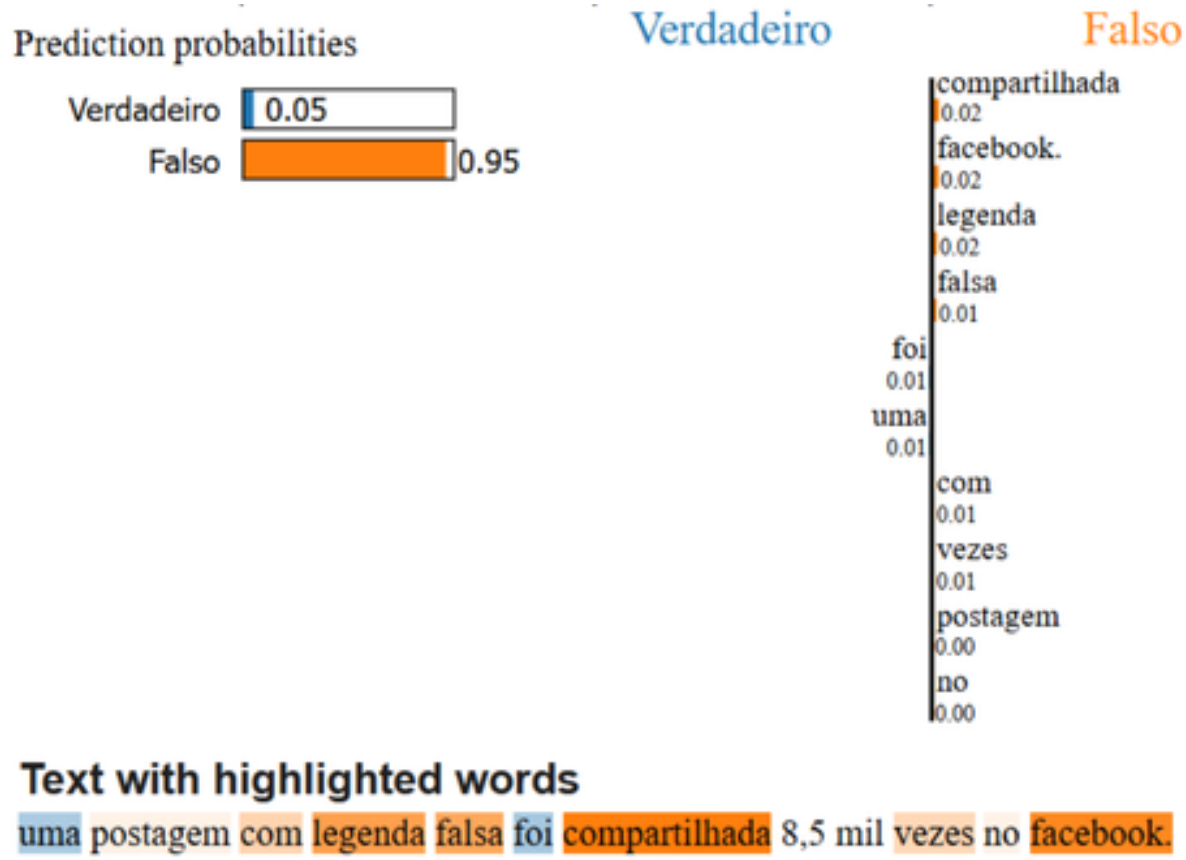
temente utilizam celebridades e eventos de entretenimento como vetores para narrativas polarizadas, explorando figuras públicas que geram alto engajamento emocional, diferentemente da pauta política institucional (“ministério”, “decreto”).

A Figura 20 exibe a interpretação do Modelo Pré-treinado para a amostra falsa. O texto é extremamente conciso e descreve a circulação de um conteúdo enganoso em rede social.

A análise demonstra que o modelo pré-treinado utilizou marcadores explícitos de viralização e descrédito para classificar a amostra:

- Rastreamento de viralização (“facebook”, “compartilhada”): Estes termos referem-se diretamente ao mecanismo de distribuição de conteúdo nas redes sociais. A sua forte influência na decisão de falsidade sugere que narrativas focadas na métrica de compartilhamento ou na plataforma de origem são características estruturais predominantes em amostras de desinformação ou em seus resumos de checagem.

Figura 20 – Explicação gerada pelo LIME do modelo pré-treinado no Drift 2 - Falso



Fonte: Elaborado pelo autor.

- Qualificadores de veracidade (“falsa”, “legenda”): A presença do adjetivo “falsa” e do substantivo “legenda” atua como um marcador direto. Diferente de notícias reais que relatam fatos, este vocabulário descreve a natureza de uma postagem, indicando que o texto pertence a um contexto de denúncia ou de identificação de fraude. A alta relevância desses termos pode indicar que um texto ou uma palavra associados ao sentido negativo é um preditor para a classe “Falso”.

É importante ressaltar a ambiguidade a esta amostra específica. O texto, embora classificado como “Falso” no conjunto de dados, é tecnicamente uma afirmação verdadeira que reporta a existência de uma fraude (“uma postagem... foi compartilhada”). Esse comportamento evidencia uma zona de confusão no conjunto de dados, onde resumos de checagem de fatos podem ter herdado o rótulo do conteúdo que denunciam.

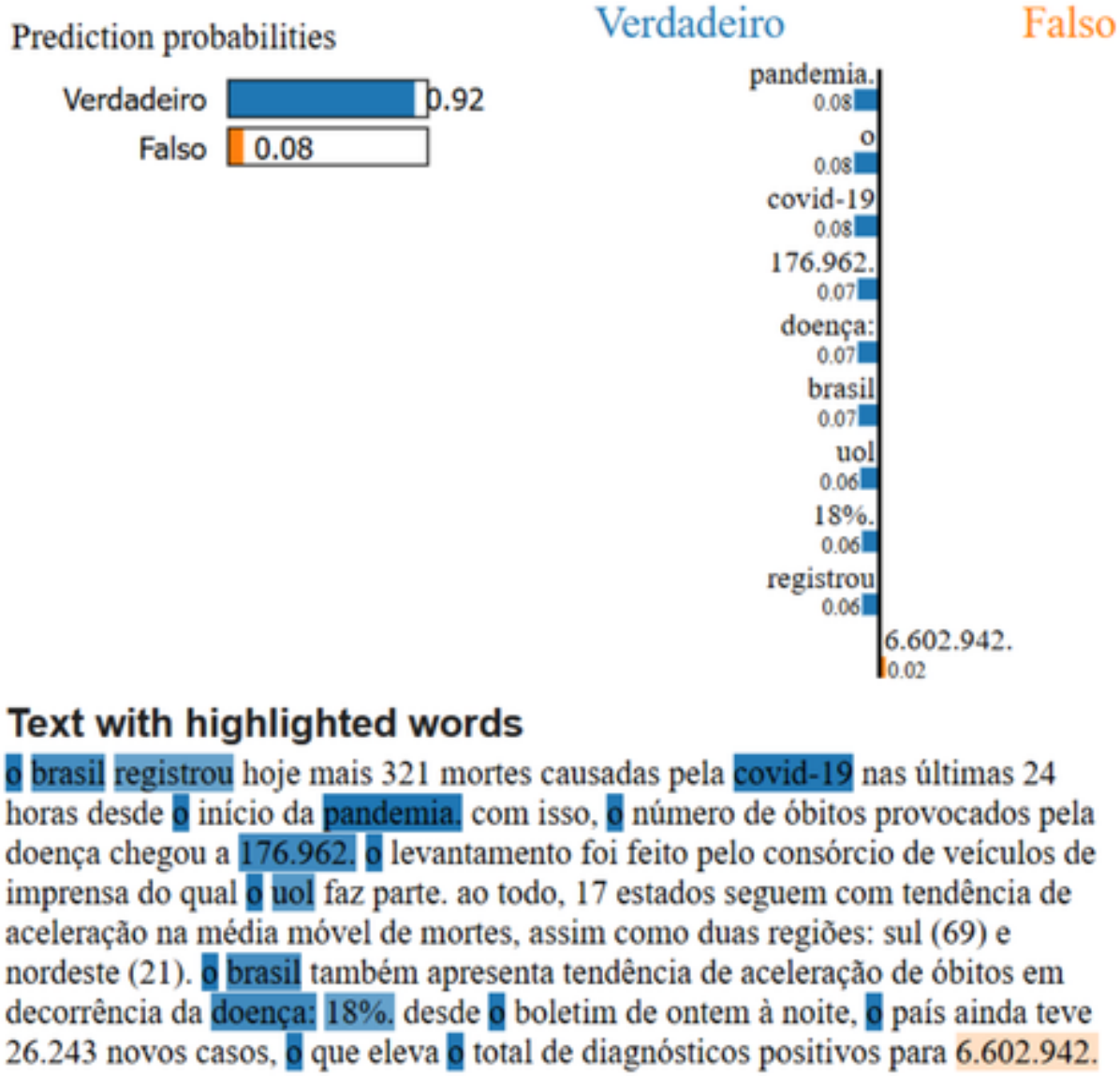
5.3.3 Análise do Drift 3

Por fim, a terceira janela de análise, denominada *Drift* 3, corresponde ao período de estabilização identificado em dezembro de 2020.

5.3.3.1 Notícias Verdadeiras

A Figura 21 exibe a explicação do Modelo de Referência para a amostra verdadeira. O texto é um boletim epidemiológico gerado pelo consórcio de veículos de imprensa, uma iniciativa criada para consolidar dados estaduais durante a crise de transparência dos dados federais no final de 2020.

Figura 21 – Explicação gerada pelo LIME do modelo de referência no Drift 3 - Verdadeiro



Fonte: Elaborado pelo autor.

- “uol”: Um veículo de notícias recebeu pontuação positiva. A citação explícita da organização responsável pela apuração ou veiculação da notícia é um indicativo de autoria definida, característica que diferencia o texto profissional de conteúdos anônimos ou sem fonte clara.

- “brasil”: Referências a países recebem pontuação positiva. A menção explícita à localização geográfica delimita o escopo do evento noticiado, sendo uma prática padrão em reportagens que informam dados nacionais.

- Precisão numérica (“176.962”, “18%”): A presença de números exatos e taxas percentuais atua como marcador de especificidade. Dados quantitativos precisos sugerem que o texto é fundamentado em levantamentos técnicos ou estatísticas oficiais.

- “o”: O artigo definido recebeu peso elevado. Assim como observado em análises anteriores, a presença frequente de artigos auxilia na manutenção da coesão textual e da estrutura gramatical normativa.

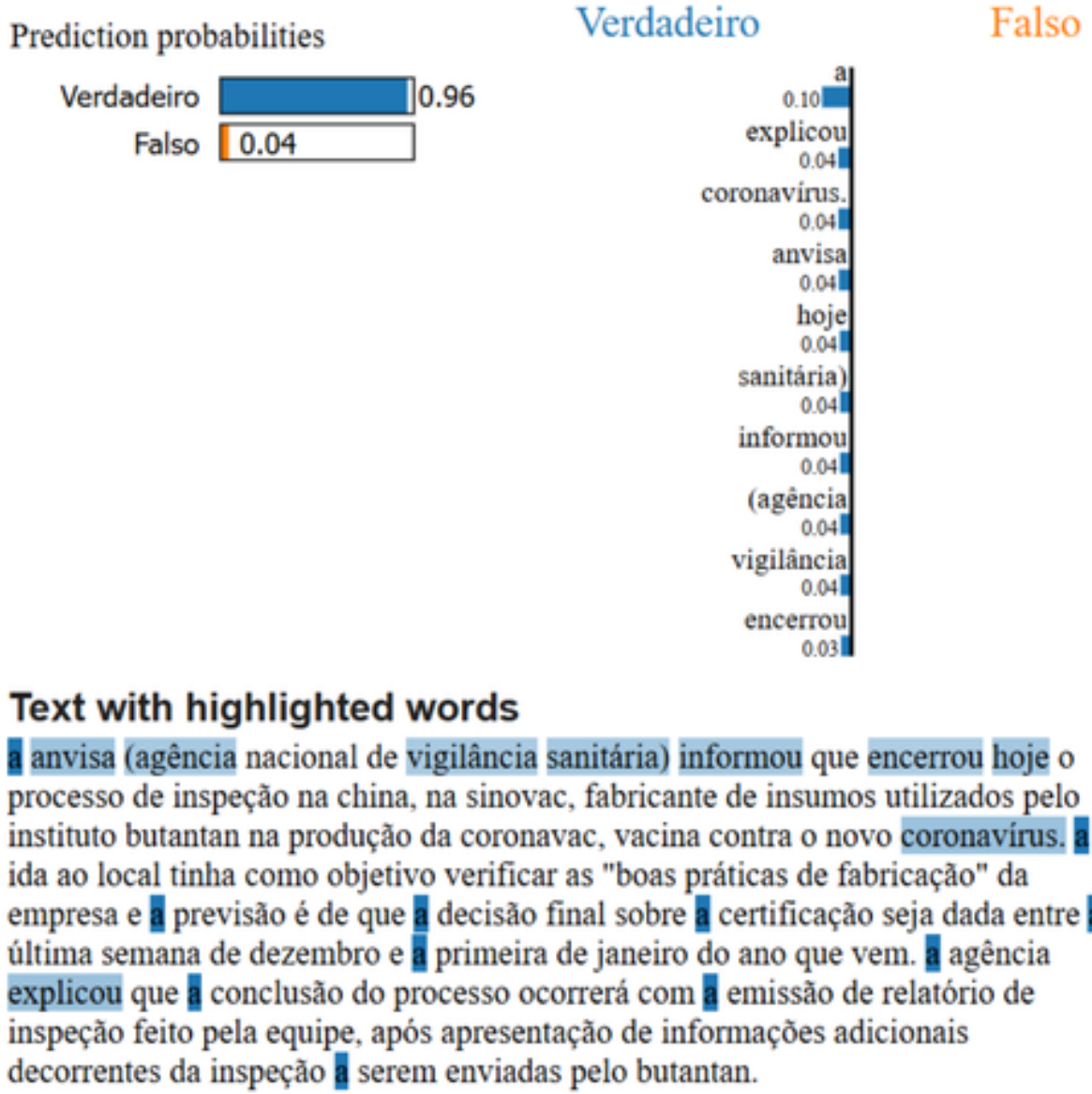
Já a Figura 22 exibe a interpretação do Modelo Pré-treinado para a amostra verdadeira. O texto aborda o processo regulatório de inspeção de vacinas na China pela autoridade sanitária brasileira, um tema técnico que dominou o noticiário no final de 2020.

A análise revela que o modelo pré-treinado combinou a detecção de entidades regulatórias com marcadores sintáticos para validar a notícia:

- Autoridade institucional e termos regulatórios (“anvisa”, “agência”, “vigilância”, “sanitária”): Este conjunto de palavras define a entidade responsável pela validação científica no país. A relevância positiva atribuída a cada parte do nome oficial da instituição sugere que a menção a órgãos reguladores atua como um forte atestado de legitimidade. Textos que detalham processos burocráticos ou técnicos de agências oficiais tendem a ser verídicos, contrastando com boatos que geralmente atacam essas instituições sem utilizar a nomenclatura técnica correta.

- Verbos de atribuição e processo (“informou”, “explicou”, “encerrou”): Estes verbos descrevem ações oficiais e comunicados. O uso de "informou"e "explicou"é típico do jornalismo declaratório, que repassa à população as decisões tomadas pelas autoridades. A presença

Figura 22 – Explicação gerada pelo LIME do modelo pré-treinado no Drift 3 - Verdadeiro



Fonte: Elaborado pelo autor.

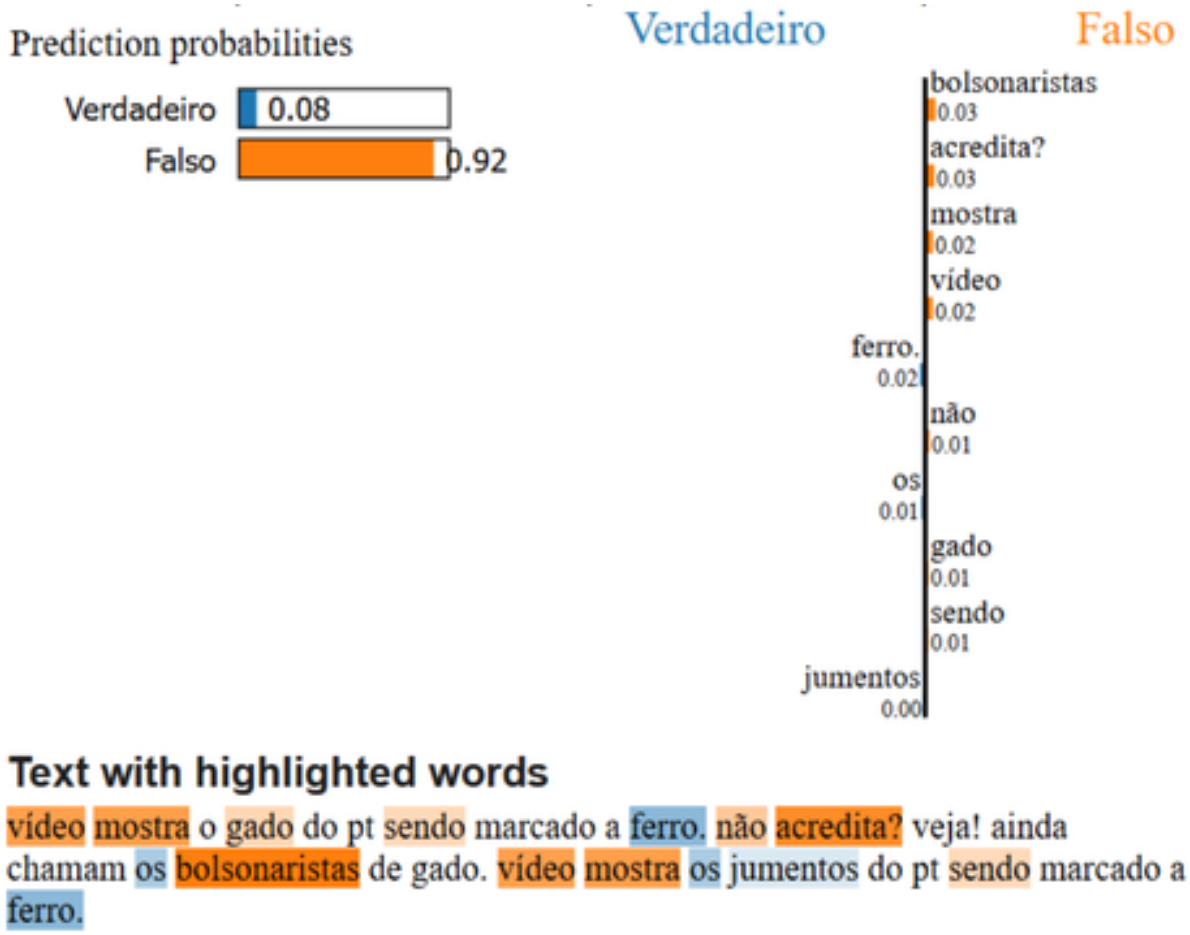
desses termos indica um texto pautado na transparência de atos administrativos, característica central da cobertura de políticas públicas de saúde.

- “a”: Neste exemplo, é possível notar que uma palavra comum que compõe a língua portuguesa foi destacada com um peso positivo. O peso positivo dessa palavra por si só não fornece muitas informações sobre a classificação, mas isso pode nos levar a considerar que notícias reais são melhor escritas e que, em muitas notícias falsas, essas palavras básicas consideradas *stop word* são escassas na estrutura do texto. A presença dessa palavra com peso positivo também pode ser explicada pelo fato de ser considerada uma palavra “neutra” em termos de conteúdo informativo.

5.3.3.2 Notícias Falsas

A Figura 23 exibe a explicação do Modelo de Referência para a amostra falsa. O texto é um exemplo claro de conteúdo inflamatório e polarizado, utilizando termos pejorativos para desumanizar grupos políticos adversários, uma tática comum em correntes de desinformação disseminadas em aplicativos de mensagens.

Figura 23 – Explicação gerada pelo LIME do modelo de referência no Drift 3 - Falso



Fonte: Elaborado pelo autor.

A análise dos pesos indica que o classificador associou a linguagem de conflito e o apelo visual à categoria falsa:

- "bolsonaristas": O termo de maior peso negativo foi a referência explícita a um grupo político específico. A presença dessa palavra, especialmente em um contexto semântico agressivo, sinaliza que o texto se desvia da neutralidade jornalística para adentrar no

campo da disputa ideológica e do partidarismo, características frequentes em conteúdos fabricados para gerar engajamento emocional negativo.

- *Clickbait* e apelo visual (“acredita?”, “vídeo”, “mostra”): Este conjunto de palavras compõe a estrutura clássica de isca de cliques. A pergunta retórica (“não acredita?”) combinada com a promessa de uma prova visual sensacionalista (“vídeo mostra”) constitui um padrão retórico voltado para induzir o usuário ao clique ou ao compartilhamento imediato, contornando a análise crítica do conteúdo.

- Termos agrícolas em contexto político (“gado”, “ferro”): Curiosamente, palavras como “ferro” e “gado” receberam pesos positivos (azul) ou neutros, apesar de serem usadas pejorativamente no texto. Isso sugere que o modelo de referência, treinado em notícias gerais, pode ter dificuldade em identificar a ironia ou a metáfora, interpretando esses termos em seu sentido literal, o que reforça a importância das *features* de polarização (“bolsonaristas”) e sintaxe de viralização para garantir a classificação correta de falsidade.

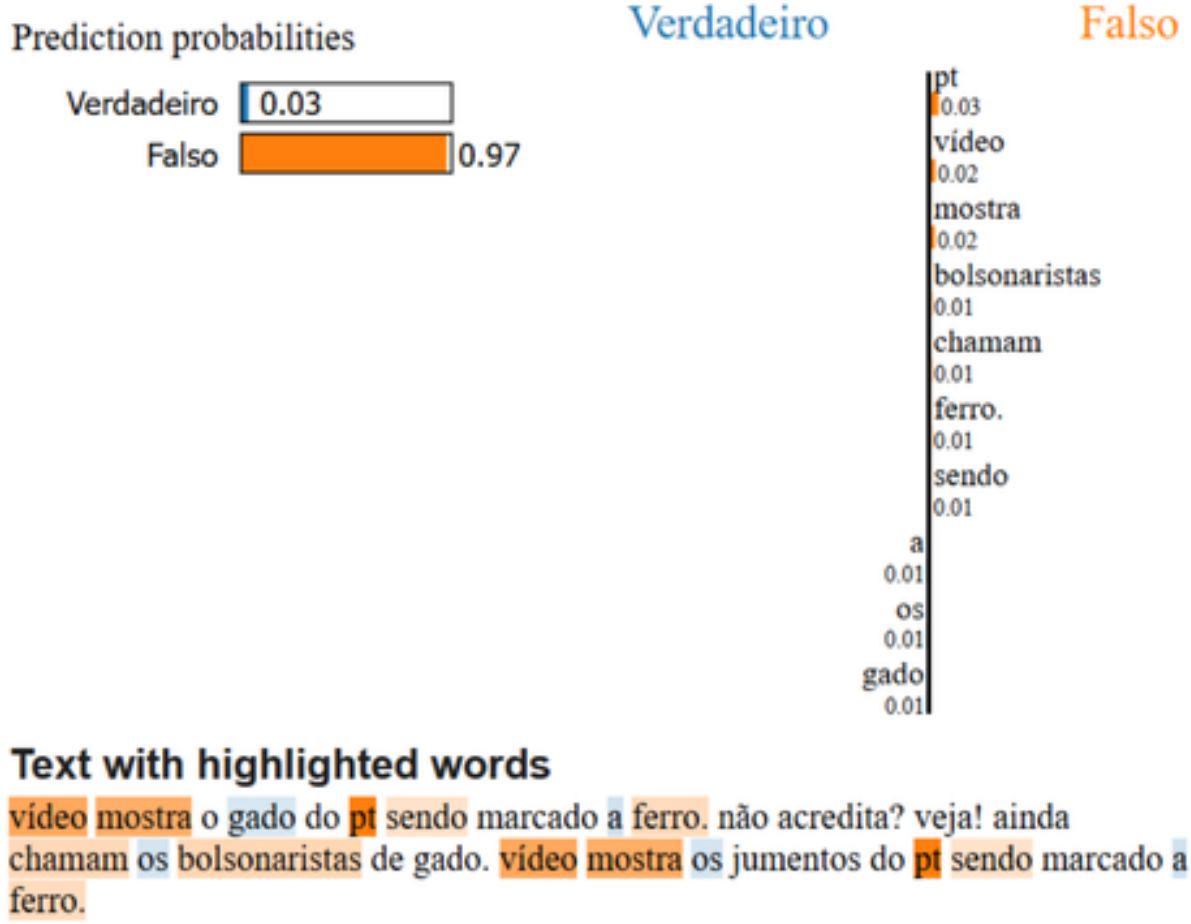
Vale destacar que essa é mais uma amostra proveniente de uma notícia de checagem de fatos. Frequentemente, esses artigos reproduzem a manchete original do boato ou o texto da postagem enganosa para contextualizar o desmentido. Isso justifica a presença de uma linguagem tão direta e ofensiva no conjunto de dados: o que o modelo está analisando, neste caso, é a transcrição fiel da desinformação que serviu de alvo para a verificação jornalística.

A Figura 24 exibe a interpretação do Modelo Pré-treinado para a mesma amostra analisada anteriormente. A comparação entre os dois modelos revela nuances interessantes: embora ambos tenham classificado o texto corretamente como falso, o modelo pré-treinado atribuiu pesos distintos aos elementos de polarização.

A análise destaca a sensibilidade do modelo a siglas partidárias e à linguagem imperativa:

- “pt”: Diferentemente do modelo de referência, que destacou o termo “bolsonaristas”, o modelo pré-treinado atribuiu o maior peso de falsidade à sigla “pt”. Isso sugere que o classificador aprendeu uma forte correlação entre a menção a este partido específico e a classe falsa dentro do conjunto de dados, possivelmente devido à alta frequência da sigla

Figura 24 – Explicação gerada pelo LIME do modelo pré-treinado no Drift 3 - Falso



Fonte: Elaborado pelo autor.

em correntes de desinformação política.

- Imperativo de prova visual (“vídeo”, “mostra”, “veja!”): O modelo reforçou a importância da chamada para ação visual. Termos que incitam o usuário a consumir uma suposta prova material atuam como gatilhos de sensacionalismo. A recorrência desses termos indica que a estrutura de "denúncia com vídeo"é um formato padrão detectado pelo modelo em conteúdos enganosos.

- Limitação semântica em metáforas (“gado”): Um ponto crucial de concordância entre os dois modelos é a incapacidade de detectar a ofensa na palavra “gado”. O termo recebeu peso positivo ou neutro, indicando que o modelo tende a interpretar a palavra em seu sentido literal, associando-a a contextos rurais verdadeiros, e falha em captar a metáfora utilizada no discurso de ódio político.

5.3.4 Discussão dos Padrões Identificados

Enquanto as métricas quantitativas apresentadas anteriormente evidenciaram a variação de desempenho decorrente do pré-treinamento via MLM, a análise qualitativa com o LIME permitiu observar os critérios de decisão adotados pelos modelos e a evolução dos temas abordados. Independentemente do período temporal, observou-se a recorrência de comportamentos que revelam heurísticas e vieses específicos. A seguir, discutem-se os padrões predominantes observados durante os experimentos.

5.3.4.1 Dinâmica Temática e Validação dos Drifts

A inspeção visual das explicações geradas pelo LIME permitiu corroborar qualitativamente os pontos de mudança estatística identificados por Wanderley *et al.* (2025). A análise léxica revelou que as alterações no vocabulário considerado relevante pelo modelo refletem diretamente a evolução do debate público sobre a pandemia.

Ao longo dos três pontos de mudança analisados, observou-se que as notícias verdadeiras mantiveram uma notável consistência temática, focadas predominantemente em termos do contexto sanitário, como “coronavírus”, “vacina”, “saúde” e dados estatísticos. Isso indica que o jornalismo profissional manteve uma pauta técnica e informativa durante todo o ano de 2020, independentemente das flutuações externas do debate público.

Por outro lado, as notícias falsas apresentaram uma mutação temática acelerada. No primeiro *drift*, a desinformação compartilhava do mesmo vocabulário sanitário das notícias reais, focando em curas falsas e origem do vírus. Contudo, a partir do segundo *drift*, notou-se uma divergência clara: enquanto a classe verdadeira permaneceu ancorada no domínio da saúde, a classe falsa migrou para um contexto majoritariamente político. Esse padrão de politização do discurso consolidou-se no terceiro *drift*, onde, embora o tema “vacina” fosse central, a abordagem nas notícias falsas era instrumentalizada para a polarização política, distanciando-se da abordagem regulatória e científica observada nas notícias verdadeiras.

5.3.4.2 Problema de rotulagem

Um dos achados mais críticos desta análise foi a tendência dos modelos em utilizar vocabulário de checagem de fatos como preditor da classe “Falso”. Observou-se que diversas amostras rotuladas como falsas no conjunto de dados não eram o texto original da desinformação,

mas sim o resumo da checagem (ex: “uma postagem falsa foi compartilhada”).

O LIME revelou que o modelo aprendeu a associar palavras que descrevem a própria notícia ou sua circulação, como “legenda”, “imagem”, “boato”, “compartilhada” à falsidade. Isso gera um paradoxo: o classificador acerta a predição, mas pelos motivos errados. Ele não está detectando a característica enganosa da narrativa, mas sim a estrutura textual de uma denúncia. Esse comportamento evidencia um ruído na construção do conjunto de dados e sugere que o modelo pode ter dificuldade em distinguir uma notícia falsa de uma notícia verdadeira que reporta a existência daquela notícia falsa.

5.3.4.3 Sintaxe e Stopwords

Ao contrário das abordagens tradicionais de Processamento de Linguagem Natural, que frequentemente removem *stopwords* (artigos, preposições, conectivos), a análise com LIME demonstrou que essas palavras desempenham um papel fundamental na decisão do BERTimbau, especialmente para a classe “Verdadeira”.

Termos como “o”, “a”, “de” e “em” receberam consistentemente pesos positivos elevados. A hipótese levantada é que o modelo utiliza a densidade de conectivos e a correção gramatical como uma variável para a credibilidade jornalística. Notícias profissionais tendem a seguir a norma culta, com uso rigoroso de regência e artigos definidos para manter a coesão textual. Em contraste, a desinformação analisada frequentemente adota um estilo telegráfico, imperativo ou coloquial, com supressão de artigos para acelerar a leitura (ex: “Vídeo mostra gado do PT”, em vez de “O vídeo mostra o gado...”). Portanto, os resultados sugerem que a complexidade sintática atua como um preditor relevante, levantando a hipótese de que o modelo aprendeu a correlacionar a formalidade da escrita com a veracidade da informação.

5.3.4.4 Vieses de Entidades e Limitações Semânticas

Por fim, a explicabilidade expôs a sensibilidade dos modelos a entidades nomeadas específicas e sua dificuldade com linguagem figurada.

Viés de Entidade: O modelo pré-treinado demonstrou uma forte correlação entre siglas partidárias (como “PT”) e a classe falsa, enquanto o modelo de referência focou em termos de grupos políticos (“bolsonaristas”). Isso indica que o classificador pode estar superajustado a tópicos políticos específicos que eram predominantes na base de treino, correndo o risco de classificar como falsa qualquer notícia futura que mencione essas entidades, independentemente

da veracidade do fato.

A análise da palavra “gado” revelou uma limitação semântica importante. O termo, usado pejorativamente no contexto político brasileiro para ridicularizar militantes, foi interpretado pelo modelo com peso neutro ou positivo, sugerindo uma leitura literal. Isso demonstra que, apesar da eficiência do modelo na captura de contexto, ele ainda falha em detectar ironias, sarcasmo e metáforas agressivas que são centrais no discurso de ódio e na polarização política.

6 CONCLUSÃO E TRABALHOS FUTUROS

O fenômeno da desinformação, potencializado pela dinâmica das redes sociais, exige soluções de detecção que sejam não apenas precisas, mas também auditáveis e robustas às mudanças temporais de contexto. Este trabalho investigou o impacto do pré-treinamento via MLM no desempenho e na interpretabilidade do modelo BERTimbau aplicado ao *corpus* FakeRecogna 2.0.

Sob a ótica quantitativa, os experimentos validaram a hipótese de que a especialização do modelo de linguagem ao contexto jornalístico brasileiro traz ganhos significativos. O Modelo Pré-treinado superou a abordagem de referência em todas as métricas de teste. Esses resultados indicam que a exposição prévia ao vocabulário não supervisionado refinou as representações vetoriais do BERTimbau, tornando-o mais sensível às nuances que distinguem notícias falsas e reduzindo a taxa de falsos negativos. Conclui-se que a especialização no domínio, por meio do pré-treinamento via MLM, foi determinante para discernir nuances textuais específicas, mostrando-se superior ao uso de modelos genéricos puramente como extratores de características.

No âmbito qualitativo, a aplicação da técnica LIME permitiu uma auditoria profunda do comportamento dos classificadores, colaborando visualmente com os pontos de mudança estatística identificados na literatura. A análise revelou que, ao longo dos três pontos de mudança analisados, as notícias verdadeiras mantiveram uma notável consistência temática, focadas predominantemente em termos do contexto sanitário e dados estatísticos, o que indica que o jornalismo profissional preservou uma pauta técnica e informativa durante todo o período. Em contrapartida, as notícias falsas apresentaram uma mutação temática acelerada, migrando de um vocabulário inicialmente sanitário para um contexto majoritariamente político e polarizado a partir do segundo *drift*.

A interpretabilidade também revelou nuances linguísticas importantes, como o fato de a sintaxe operar como um forte preditor de veracidade para o modelo, que associa a presença de conectivos e a norma culta à credibilidade. Por outro lado, a estrutura de viralização consolidou-se como um marcador estrutural de falsidade. Entretanto, a pesquisa expôs limitações, como a dificuldade do modelo em interpretar ironias e metáforas agressivas, além de um ruído crítico no conjunto de dados, onde textos de checagem de fatos foram classificados corretamente como falsos baseando-se na estrutura da denúncia, e não no boato em si.

Como trabalhos futuros, vislumbram-se diversas frentes de expansão para consolidar e ampliar os achados desta pesquisa.

No âmbito de arquiteturas e estratégias de treinamento, sugere-se a aplicação da metodologia proposta a outros modelos de linguagem baseados em Transformers (como RoBERTa, DistilBERT ou DeBERTa) para verificar se os benefícios do pré-treinamento via MLM são consistentes através de diferentes arquiteturas. Adicionalmente, propõe-se investigar o impacto do *fine-tuning* completo (descongelamento de todas as camadas) em comparação à abordagem de extração de características aqui adotada, visando determinar se o ajuste integral dos parâmetros justifica o custo computacional com ganhos adicionais de performance.

Em relação à generalização e dados, recomenda-se a execução de protocolos de avaliação cruzada (*cross-dataset*), onde o pré-treinamento é realizado em um conjunto de dados e o ajuste fino final em outro, distinto. Essa abordagem permitiria testar a robustez da adaptação de domínio e a capacidade do modelo de transferir o conhecimento aprendido para novos contextos de desinformação, mitigando o risco de sobreajuste a características específicas de um único *corpus*.

Por fim, na vertente da interpretabilidade, planeja-se superar as limitações de métodos baseados em perturbação local através da adoção de um conjunto diversificado de técnicas de XAI. Sugere-se a exploração de: Visualização de Atenção, para mapear quais tokens recebem maior foco nas camadas internas do modelo; Mapas de Saliência baseados em gradientes, para evidenciar matematicamente os termos mais sensíveis à decisão de classificação; e abordagens híbridas utilizando *Retrieval-Augmented Generation* (RAG), visando fornecer explicações contrastivas fundamentadas em fontes de verificação externas, elevando o patamar de confiabilidade do sistema.

REFERÊNCIAS

ALVES, M. A.; ANDRADE, O. Da "caixa-preta"à "caixa de vidro": o uso da explainable artificial intelligence (xai) para reduzir a opacidade e enfrentar o enviesamento em modelos algorítmicos. **Direito Público**, v. 18, n. 100, jan. 2022.

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.

CAMBRIA, E.; WHITE, B. Jumping nlp curves: A review of natural language processing research. **IEEE Computational intelligence magazine**, IEEE, v. 9, n. 2, p. 48–57, 2014.

CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de linguagem natural**: conceitos, técnicas e aplicações em português. 3. ed. BPLN, 2024. ISBN 978-65-01-20581-6. Disponível em: https://brasileiraspln.com/livro-pln/3a-edicao/. Acesso em: 12 jun.2025.

CERRI, R. Aprendizado de mÁquina: Breve introdução e aplicações1. **Cadernos de Ciência & Tecnologia**, v. 34, n. 3, p. 297–313, 2017.

CHOI, R. Y.; COYNER, A. S.; KALPATHY-CRAMER, J.; CHIANG, M. F.; CAMPBELL, J. P. Introduction to machine learning, neural networks, and deep learning. **Translational vision science & technology**, The Association for Research in Vision and Ophthalmology, v. 9, n. 2, p. 14–14, 2020.

CHOLLET, F. **Deep learning with Python**. [*S. l.*]: simon and schuster, 2021.

DANESH, A. S.; REZANEJAD, A. Truth is all you need: Enhancing fake news detection with interpretable language models. **Preprints**, Preprints, November 2025. Disponível em: https://doi.org/10.20944/preprints202511.0029.v1.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: BURSTEIN, J.; DORAN, C.; SOLORIO, T. (Ed.). **Proceedings [...]**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: https://aclanthology.org/N19-1423/. Acesso em: 14 mai.2025.

FLECK, L.; TAVARES, M. H. F.; EYNG, E.; HELMANN, A. C.; ANDRADE, M. A. d. M. Redes neurais artificiais: Princípios básicos. **Revista Eletrônica Científica Inovação e Tecnologia**, v. 1, n. 13, p. 47–57, 2016.

GARCIA, G. L.; PAIOLA, P. H.; JODAS, D. S.; SUGI, L. A.; PAPA, J. P. Text summarization and temporal learning models applied to Portuguese fake news detection in a novel Brazilian corpus dataset. In: GAMALLO, P.; CLARO, D.; TEIXEIRA, A.; REAL, L.; GARCIA, M.; OLIVEIRA, H. G.; AMARO, R. (Ed.). **Anais [...]**. Association for Computational Lingustics, 2024. p. 86–96. Disponível em: https://aclanthology.org/2024.propor-1.9/. Acesso em: 4 jun.2025.

GéRON, A. **Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow:**: concepts, tools, and techniques to build intelligent systems. Second edition. Sebastopol, CA: O'Reilly, 2019. ISBN 9781492032618. Disponível em: https://ebookcentral.proquest.com/lib/CM/detail.action?docID=5892320. Acesso em: 1 abr.2025.

GUNNING, D.; AHA, D. Darpa's explainable artificial intelligence (xai) program. **AI magazine**, v. 40, n. 2, p. 44–58, 2019.

HAMED, S. K.; AZIZ, M. J. A.; YAAKUB, M. R. A review of fake news detection approaches: A critical analysis of relevant studies. **Heliyon**, v. 9, n. 10, 2023.

HAYKIN, S. **Redes neurais:**: princípios e prática. 2. ed. Porto Alegre, RS: Bookman Editora, 2001. Tradução de Paulo Martins Engel. ISBN 9788573077186.

KANNEGANTI, D. **A New cross-domain strategy based XAI models for fake news detection**. 2023. Disponível em: https://arxiv.org/abs/2302.02122. Acesso em: 6 jan.2026.

KIM, B.; XIONG, A.; LEE, D.; HAN, K. A systematic review on fake news research through the lens of news creation and consumption: Research efforts, challenges, and future directions. **PLOS ONE**, v. 16, n. 12, p. 1–28, 2021.

KüHL, N.; SCHEMMER, M.; GOUTIER, M.; SATZGER, G. Artificial intelligence and machine learning. **Electronic Markets**, Springer, v. 32, n. 4, p. 2235–2244, 2022.

LAZER, D. M.; BAUM, M. A.; BENKLER, Y.; BERINSKY, A. J.; GREENHILL, K. M.; MENCZER, F. *et al.* The science of fake news. **Science**, v. 359, n. 6380, p. 1094–1096, 2018.

MENDONÇA, R. F.; FREITAS, V. G.; AGGIO, C. d. O.; SANTOS, N. F. d. Fake news e o repertório contemporâneo de ação política. **Dados**, v. 66, n. 2, p. e20200213, 2022.

MISHIMA, K.; YAMANA, H. A survey on explainable fake news detection. **IEICE TRANSACTIONS on Information and Systems**, The Institute of Electronics, Information and Communication Engineers, v. 105, n. 7, p. 1249–1257, 2022.

MOLNAR, C. **Interpretable Machine Learning**: A guide for making black box models explainable. 3. ed. [*S. n.*], 2025. ISBN 978-3-911578-03-5. Disponível em: https://christophm.github.io/interpretable-ml-book. Acesso em: 31 jul.2025.

MOURATIDIS, D.; KANAVOS, A.; KERMANIDIS, K. From misinformation to insight: Machine learning strategies for fake news detection. **Information**, v. 16, n. 3, 2025.

RAHMADHANI, B.; PURWONO, P.; KURNIAWAN, S. D. Understanding transformers: A comprehensive review. **Journal of Advanced Health Informatics Research**, v. 2, n. 2, p. 85–94, 2024.

RAUBER, T. W. **Redes neurais artificiais**. Universidade Federal do Espírito Santo: [*S. n.*], 2005. v. 29. 39 p.

REIS, J. C.; CORREIA, A.; MURAI, F.; VELOSO, A.; BENEVENUTO, F. Supervised learning for fake news detection. **IEEE Intelligent Systems**, IEEE, v. 34, n. 2, p. 76–81, 2019.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?"explaining the predictions of any classifier. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22. **Proceedings [...]**. [*S. l.*]: Association for Computing Machinery, 2016. p. 1135–1144.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. **Nature machine intelligence**, Nature Publishing Group UK London, v. 1, n. 5, p. 206–215, 2019.

SCHUSTER, T.; SCHUSTER, R.; SHAH, D. J.; BARZILAY, R. The limitations of stylometry for detecting machine-generated fake news. **Computational Linguistics**, v. 46, n. 2, p. 499–510, 06 2020. ISSN 0891-2017. Disponível em: https://doi.org/10.1162/coli_a_00380. Acesso em: 19 jan.2026.

SHU, K.; SLIVA, A.; WANG, S.; TANG, J.; LIU, H. Fake news detection on social media: A data mining perspective. **ACM SIGKDD Explorations Newsletter**, ACM, v. 19, n. 1, p. 22–36, 2017.

SHU, K.; WANG, S.; LEE, D.; LIU, H. Mining disinformation and fake news: Concepts, methods, and recent advancements. In: SHU, K.; WANG, S.; LEE, D.; LIU, H. (Ed.). **Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities**. Cham: Springer International Publishing, 2020. p. 1–19. ISBN 978-3-030-42699-6. Disponível em: https://doi.org/10.1007/978-3-030-42699-6_1. Acesso em: 8 abr.2025.

SILVA, F. d. C.; FEITOSA, R. M.; BATISTA, L. A.; SANTANA, A. M. Análise comparativa de métodos de explicabilidade da inteligência artificial no cenário educacional: um estudo de caso sobre evasão. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE). **Anais [...]**. [*S. l.*]: SBC, 2024. p. 2968–2977.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS, 9., Rio Grande, Brazil. **Anais [...]**. [*S. l.*]: Springer-Verlag, 2020. p. 403–417.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. USA: Addison-Wesley Longman Publishing Co., Inc., 2005. ISBN 0321321367.

THAKAR, H.; BHATT, B. Fake news detection: recent trends and challenges. **Social Network Analysis and Mining**, v. 14, n. 1, p. 176, 2024.

TUNSTALL, L.; WERRA, L. V.; WOLF, T. **Natural language processing with transformers**. [*S. l.*]: "O'Reilly Media, Inc.", 2022.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017.

VICENTINI, J. U. **Comparando técnicas de explicabilidade sobre modelos de linguagem: um estudo de caso na detecção de notícias falsas**. Dissertação (Mestrado em Ciência da Computação) – Universidade Estadual Paulista (Unesp), São José do Rio Preto, jul 29 2024. Disponível em: https://hdl.handle.net/11449/257181. Acesso em: 18 nov.2025.

VOSOUGHI, S.; ROY, D.; ARAL, S. The spread of true and false news online. **Science**, v. 359, n. 6380, p. 1146–1151, 2018.

WANDERLEY, M.; FERRAZ, L.; ALMEIDA, T.; SILVA, R. A moving target: Detecting concept drift in brazilian portuguese fake news. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 16. **Anais [...]**. Porto Alegre, RS, Brasil: SBC, 2025. p. 490–501. ISSN 0000-0000. Disponível em: https://sol.sbc.org.br/index.php/stil/article/view/37849. Acesso em: 30 out.2025.

ZHOU, X.; ELFARDY, H.; CHRISTODOULOPOULOS, C.; BUTLER, T.; BANSAL, M. **Hidden Biases in Unreliable News Detection Datasets**. 2021. Disponível em: https://arxiv.org/abs/2104.10130. Acesso em: 15 jan.2026.