Avaliação da Generalização de Modelos BERTimbau na Detecção de Desinformação em Português com Análise Qualitativa de Dados

Pedro Henrique Ramos
Faculdade de Computação (FACOM)
Universidade Federal de Mato Grosso do Sul (UFMS)
Campo Grande, Brasil
Email: rohp2002@gmail.com

Orientador: Prof. Dr. Bruno Magalhães Nogueira
Faculdade de Computação (FACOM)
Universidade Federal de Mato Grosso do Sul (UFMS)
Campo Grande, Brasil

Maio de 2026

***Resumo*—A crescente disseminação de informações na internet tem intensificado a propagação de notícias falsas, tornando a detecção automática de fake news um problema relevante. Nesse contexto, modelos de linguagem pré-treinados têm apresentado desempenho significativo em tarefas de processamento de linguagem natural, incluindo a classificação de textos em português.**

Este trabalho avalia o desempenho do modelo *BERTimbau* na classificação de notícias como verdadeiras ou falsas, utilizando múltiplas bases de dados com características distintas. Para garantir uma avaliação equilibrada, os conjuntos de dados foram previamente balanceados em relação às classes. Os experimentos foram conduzidos com o objetivo de analisar não apenas o desempenho do modelo no domínio de treinamento, mas também sua capacidade de generalização entre diferentes bases.

Adicionalmente, foi incorporada uma etapa de análise qualitativa baseada em modelo de linguagem de grande porte, com o objetivo de examinar padrões textuais, diferenças entre os conjuntos de dados e aspectos recorrentes associados aos acertos e erros do classificador. Essa análise não teve como objetivo explicar formalmente os mecanismos internos do modelo, mas fornecer uma leitura complementar sobre características linguísticas e contextuais presentes nas bases avaliadas.

***Palavras-chave*—Fake news, BERTimbau, Processamento de Linguagem Natural, Análise Qualitativa, Dataset Shift, Desinformação.**

I. Introdução

O aumento exponencial da produção e disseminação de informações na internet tem contribuído significativamente para a propagação de desinformação, especialmente na forma de notícias falsas. Esse fenômeno representa um desafio relevante para a sociedade, impulsionando o desenvolvimento de métodos automáticos para detecção de conteúdos enganosos. Nesse cenário, técnicas de Processamento de Linguagem Natural (PLN) têm sido amplamente utilizadas para a identificação de fake news.

Nos últimos anos, modelos de linguagem pré-treinados têm se destacado pelo alto desempenho em tarefas de classificação de texto. No contexto da língua portuguesa, o modelo *BERTimbau* apresenta relevância por ser treinado especificamente para capturar padrões linguísticos do português brasileiro [1]. A partir de técnicas de ajuste fino (*fine-tuning*), modelos baseados na arquitetura BERT podem ser adaptados para tarefas específicas, como a classificação de notícias em categorias de verdadeiro ou falso [2].

Apesar dos avanços, ainda existem desafios importantes relacionados à robustez e à capacidade de generalização desses modelos. Em muitos casos, modelos treinados em uma base de dados específica apresentam queda significativa de desempenho quando aplicados a dados provenientes de outras fontes. Essa limitação é especialmente crítica na detecção de fake news, uma vez que o conteúdo pode variar amplamente em termos de estilo, domínio, contexto e distribuição temática.

Outro aspecto relevante diz respeito à interpretabilidade dos modelos. Modelos de linguagem pré-treinados são frequentemente tratados como sistemas de caixa-preta, dificultando a compreensão dos fatores que influenciam suas decisões. Nesse sentido, a área de Inteligência Artificial Explicável (XAI) busca desenvolver métodos capazes de tornar as decisões dos modelos mais compreensíveis para usuários e pesquisadores [3]. Neste trabalho, entretanto, a etapa baseada em modelo de linguagem foi tratada como análise qualitativa auxiliar, voltada à investigação de padrões textuais e contextuais associados aos resultados, e não como um método formal de explicabilidade das decisões internas do BERTimbau.

Diante desse contexto, este trabalho tem como objetivo avaliar o desempenho de um modelo de linguagem pré-treinado na classificação de notícias em português, utilizando múltiplas bases de dados. O estudo busca analisar não apenas métricas tradicionais de desempenho, mas também a capacidade de generalização do modelo em diferentes cenários. Além disso, é incorporada uma etapa de análise qualitativa baseada em modelo de linguagem de grande porte, com o intuito de examinar padrões textuais, ambiguidades e características recorrentes nas bases de dados e nos casos de acerto e erro do classificador.

Como contribuições, destacam-se: (i) a avaliação do modelo em diferentes bases de dados; (ii) a análise de sua capacidade de generalização em cenários com mudança de domínio; (iii) a comparação com um classificador MLP como baseline; e

(iv) a incorporação de uma análise qualitativa auxiliar para melhor compreensão das diferenças entre os datasets e do comportamento observado nos experimentos

II. Trabalhos Relacionados

A detecção automática de fake news tem sido amplamente investigada na literatura, especialmente com o avanço de técnicas de aprendizado de máquina e Processamento de Linguagem Natural. Trabalhos anteriores exploram desde abordagens baseadas em características léxicas e estatísticas até modelos mais recentes baseados em redes neurais profundas e transformadores.

O uso de modelos baseados em BERT tornou-se uma alternativa relevante para tarefas de classificação textual, uma vez que esses modelos são capazes de representar o contexto bidirecional das palavras em uma sentença [2]. Para a língua portuguesa, o BERTimbau foi proposto como uma versão pré-treinada em português brasileiro, apresentando resultados competitivos em diferentes tarefas de PLN [1].

No contexto específico de fake news em português, bases como FakeBR e FactckBR têm sido utilizadas em pesquisas de classificação textual. O FakeBR é composto por notícias verdadeiras e falsas alinhadas em português brasileiro [4], enquanto o FactckBR reúne notícias supostamente falsas associadas a checagens e classificações [5]. Essas bases permitem o desenvolvimento e a avaliação de modelos supervisionados para detecção de desinformação.

Entretanto, um desafio recorrente é que bons resultados em uma única base de dados não garantem necessariamente bom desempenho em cenários reais. Quando há diferença entre a distribuição dos dados de treinamento e dos dados de teste, ocorre o fenômeno conhecido como *dataset shift*, que pode comprometer a capacidade de generalização dos modelos [6]. Esse problema é particularmente relevante em fake news, pois diferentes bases podem variar em estilo, tamanho dos textos, temas predominantes e critérios de rotulagem.

Além disso, a interpretabilidade dos modelos tem ganhado importância em aplicações sensíveis, como detecção de desinformação. A área de Inteligência Artificial Explicável (XAI) busca desenvolver métodos capazes de tornar as decisões dos modelos mais compreensíveis para usuários e pesquisadores [3]. Neste trabalho, entretanto, a etapa baseada em modelo de linguagem foi tratada como análise qualitativa auxiliar, voltada à investigação de padrões textuais e contextuais associados aos resultados, e não como um método formal de explicabilidade das decisões internas do BERTimbau.

III. Metodologia

Esta seção descreve os conjuntos de dados utilizados, as etapas de pré-processamento, os modelos empregados nos experimentos, bem como os métodos de treinamento, avaliação e análise qualitativa adotados neste trabalho.

A. Bases de Dados

Foram utilizadas três bases de dados distintas voltadas para classificação de notícias em português: FakeBR, LupaAFP e FactckBR. Os conjuntos de dados apresentam características textuais diferentes, permitindo avaliar não apenas o desempenho do modelo no domínio de treinamento, mas também sua capacidade de generalização em cenários com mudança de domínio.

A base FakeBR consiste em um corpus de notícias verdadeiras e falsas em português brasileiro, desenvolvido para pesquisas em detecção automática de fake news [4]. Neste trabalho, foi utilizada uma versão previamente processada da base, contendo textos já normalizados e estruturados para treinamento. [7]

A base LupaAFP foi construída a partir de conteúdos oriundos de processos de checagem jornalística realizados por agências de verificação de fatos. Diferentemente da FakeBR, essa base apresenta textos com características mais próximas de conteúdos verificativos, frequentemente mais curtos, diretos e associados a publicações checadas em ambiente digital.

A FactckBR é uma base construída a partir de registros de checagem de fatos em português, coletados por meio do schema ClaimReview, utilizado por agências de fact-checking para disponibilizar resultados estruturados de verificações. Diferentemente de bases compostas por notícias jornalísticas completas, cada registro da FactckBR corresponde a uma alegação verificada, acompanhada de metadados como endereço do artigo de checagem, data de publicação, alegação analisada, texto da checagem, título, valor de classificação e rótulo textual. A base reúne registros provenientes de agências como Aos Fatos, Lupa e Truco. Por essa razão, seus textos tendem a ser mais curtos e mais próximos do gênero textual de checagem de fatos ou alegações verificadas do que de notícias completas. [5]

As bases utilizadas possuem tamanhos distintos, tanto em número total de notícias quanto na distribuição entre as classes verdadeiro e falso. Com o objetivo de reduzir possíveis vieses relacionados ao desbalanceamento de classes, todas as bases foram balanceadas previamente, mantendo a mesma quantidade de amostras para as classes *Fake* e *True*. O balanceamento foi realizado utilizando como referência a menor quantidade disponível entre as classes de cada conjunto de dados.

Essa diferença de natureza textual é relevante para a interpretação dos resultados. Como a FactckBR é formada por alegações e registros de checagem, e não necessariamente por notícias completas, seu uso em conjunto com bases como FakeBR pode introduzir diferenças de domínio relacionadas ao tamanho dos textos, ao estilo linguístico, ao vocabulário e aos critérios de rotulagem. Assim, os resultados envolvendo a FactckBR foram interpretados com cautela, especialmente nos cenários intra-domain, cross-domain e multi-domínio.

B. Caracterização dos Datasets

Além da quantidade de amostras, foram extraídas métricas descritivas dos conjuntos de dados, incluindo média de palavras por notícia, tamanho do vocabulário e temas predominantes. A Tabela I apresenta as características consolidadas das bases utilizadas.

Tabela I
CARACTERÍSTICAS CONSOLIDADAS DOS DATASETS UTILIZADOS

Dataset	Notícias	Média Palavras	Vocabulário
FakeBR	7200	366.28	79551
LupaAFP	5440	16.94	19796
FactckBR	246	22.35	2258

Os resultados mostram diferenças relevantes entre os datasets. A base FakeBR apresenta textos substancialmente maiores e vocabulário mais amplo, enquanto LupaAFP e FactckBR apresentam textos mais curtos. Essas diferenças são importantes para a análise de generalização, pois indicam variações linguísticas e estruturais significativas entre os domínios avaliados.

No caso da FactckBR, essa diferença é intensificada pelo fato de o campo claimReviewed ter sido utilizado como texto de entrada. Assim, as instâncias dessa base representam alegações verificadas, geralmente mais curtas do que notícias completas. Essa diferença de natureza textual foi considerada na interpretação dos resultados, especialmente nos experimentos de generalização.

C. Pré-processamento

As notícias passaram por uma etapa de pré-processamento textual antes do treinamento do modelo. Essa etapa teve como objetivo reduzir ruídos presentes nos textos e padronizar os dados utilizados nos experimentos.

Entre os procedimentos realizados destacam-se:

- remoção de caracteres especiais;
- normalização textual;
- remoção de elementos irrelevantes;
- padronização estrutural do conteúdo textual;
- remoção de registros com textos ausentes.

Após o pré-processamento, as bases foram convertidas para uma estrutura comum contendo as colunas text e label. A coluna text representa o conteúdo textual utilizado como entrada dos modelos, enquanto label representa a classe binária associada ao exemplo. No caso da FactckBR, a coluna text foi formada a partir do campo claimReviewed, correspondente à alegação analisada no processo de checagem.

No caso da base FakeBR, foi utilizada uma versão previamente pré-processada disponibilizada junto ao conjunto de dados original. As demais bases passaram por etapas adicionais de preparação para adequação ao pipeline experimental utilizado.

Após o pré-processamento, os textos foram convertidos em representações numéricas por meio do tokenizador compatível com o modelo BERTimbau. O tokenizador utilizado pertence ao modelo *neuralmind/bert-base-portuguese-cased* e é responsável por dividir os textos em subpalavras (*tokens*), convertendo-as em identificadores numéricos compatíveis com o vocabulário aprendido pelo modelo durante o pré-treinamento.

D. Modelos Utilizados

O principal modelo utilizado neste trabalho foi o *BERTimbau*, especificamente a versão *neuralmind/bert-base-portuguese-cased*, baseada na arquitetura BERT e pré-treinada para a língua portuguesa [1]. O modelo foi empregado utilizando a estratégia de ajuste fino (*fine-tuning*) para a tarefa de classificação binária de notícias.

A implementação foi realizada utilizando a biblioteca Hugging Face Transformers, por meio da *Trainer API*, com PyTorch como backend computacional. A biblioteca Scikit-Learn foi utilizada para separação dos dados, geração dos relatórios de classificação e cálculo das métricas.

Além do BERTimbau, foi utilizado um classificador MLP (*Multilayer Perceptron*) como baseline. Nos experimentos com MLP, os textos foram representados por vetores TF-IDF, permitindo comparar o comportamento de uma abordagem tradicional com o modelo baseado em transformadores.

E. Treinamento e Avaliação

Os experimentos foram divididos em diferentes cenários de avaliação, permitindo analisar tanto o desempenho intra-base quanto a capacidade de generalização do modelo.

Nos testes intra-domain, o modelo foi treinado e avaliado utilizando dados provenientes da mesma base de dados, com separação entre conjuntos de treino e teste. Já nos testes cross-domain, o treinamento foi realizado em uma base específica e a avaliação conduzida em outra base distinta, permitindo investigar o impacto da mudança de domínio no desempenho do modelo.

Também foram realizados experimentos multi-domínio, nos quais duas bases foram combinadas para treinamento e a avaliação foi realizada em uma terceira base. Por fim, foi conduzido um experimento com a combinação das três bases de dados em um conjunto unificado, denominado neste trabalho como *Super Mix*.

A tokenização dos textos foi realizada utilizando truncamento e preenchimento (*padding*) para o tamanho máximo de 128 tokens. O treinamento do BERTimbau foi configurado com taxa de aprendizado de 2×10^{-5}, tamanho de lote igual a 16, decaimento de peso de 0.01 e entre 2 e 3 épocas de treinamento, conforme o cenário experimental. O número de épocas foi definido considerando limitações computacionais e tempo de processamento disponíveis durante os experimentos.

O limite de 128 tokens foi adotado por restrições computacionais e para padronizar o tamanho das entradas. Como o BERTimbau utiliza tokenização em subpalavras, esse limite não corresponde necessariamente a 128 palavras. Em bases com textos longos, como a FakeBR, essa escolha pode resultar em truncamento de parte do conteúdo, sendo considerada uma limitação experimental.

Os experimentos principais foram executados no ambiente Google Colaboratory, utilizando GPU NVIDIA T4 nos cenários de maior custo computacional.

A avaliação dos modelos foi conduzida utilizando as métricas:

- Acurácia;

- Precisão;
- Revocação;
- F1-score.

F. Análise Qualitativa com Modelos de Linguagem

Além da avaliação quantitativa, este trabalho incorporou uma etapa de análise qualitativa baseada em modelo de linguagem de grande porte.

Para essa etapa, foi utilizado o modelo GPT-4o mini como ferramenta auxiliar para examinar padrões textuais, aspectos contextuais e características recorrentes nas bases e nos casos de acerto e erro do classificador. O objetivo não foi substituir o classificador principal nem explicar formalmente seus mecanismos internos, mas complementar a análise dos resultados por meio de uma leitura qualitativa dos conteúdos avaliados.

Foram selecionadas amostras de 50 notícias ou instâncias de cada base para análise qualitativa. Em uma das etapas, o prompt solicitava a extração de três características técnicas, com até três palavras cada, associadas ao conteúdo analisado. Essa restrição foi utilizada para tornar os aspectos extraídos mais objetivos, comparáveis e adequados à análise qualitativa.

As análises geradas pelo GPT-4o mini foram utilizadas como mecanismo complementar de auditoria qualitativa dos resultados, permitindo investigar possíveis padrões semânticos, ambiguidades textuais e conflitos entre a classificação produzida pelo BERTimbau e os aspectos textuais identificados durante a interpretação.

IV. Resultados e Discussão

A. Resultados Intra-Domain

Os experimentos intra-domain foram realizados treinando e avaliando o modelo utilizando dados provenientes da mesma base de dados. O objetivo dessa etapa foi verificar a capacidade do modelo em aprender padrões específicos de cada conjunto de dados.

Nos experimentos intra-domain, foram utilizados os arquivos previamente separados em treino e teste para cada base, mantendo a avaliação dentro do mesmo domínio.

A Tabela II apresenta os resultados obtidos utilizando o modelo *BERTimbau* nos cenários intra-domain.

Tabela II
Resultados intra-domain utilizando BERTimbau

Dataset	Acurácia	F1 Macro
FakeBR	0.93	0.93
LupaAFP	0.80	0.80
FactckBR	0.65	0.65

Os resultados demonstram que o modelo apresentou desempenho elevado quando treinado e avaliado no mesmo domínio, especialmente na base FakeBR, que atingiu acurácia de 0.93. Esse comportamento sugere que o modelo foi capaz de aprender padrões linguísticos e estruturais característicos do conjunto de dados.

A base LupaAFP também apresentou desempenho satisfatório, com acurácia de 0.80, enquanto a FactckBR obteve desempenho inferior em comparação às demais bases. Esse resultado pode estar relacionado à menor quantidade de amostras disponíveis nessa base e à maior dificuldade de identificação de padrões consistentes em um conjunto reduzido.

De forma geral, os resultados intra-domain indicam que o modelo possui boa capacidade de adaptação ao domínio específico em que foi treinado. Entretanto, tais resultados não garantem necessariamente robustez em cenários com mudança de domínio, aspecto investigado nos experimentos cross-domain.

B. Resultados Cross-Domain

Os experimentos cross-domain foram conduzidos com o objetivo de avaliar a capacidade de generalização do modelo em cenários com mudança de domínio. Nesses experimentos, o treinamento foi realizado em uma base de dados específica, enquanto a avaliação foi conduzida em uma base distinta.

Nos experimentos cross-domain, os conjuntos de treino e teste de uma base foram reunidos para formar o conjunto completo de treinamento, enquanto a avaliação foi realizada sobre outra base completa. Dessa forma, no cenário FakeBR $\rightarrow$ LupaAFP, por exemplo, o modelo foi treinado com a FakeBR completa e avaliado na LupaAFP completa.

A Tabela III apresenta os resultados obtidos utilizando o modelo *BERTimbau* nos cenários cross-domain.

Tabela III
Resultados cross-domain utilizando BERTimbau

Treino $\rightarrow$ Teste	Acurácia	F1 Macro
FakeBR $\rightarrow$ LupaAFP	0.50	0.34
LupaAFP $\rightarrow$ FakeBR	0.50	0.33

Os resultados demonstram uma queda significativa de desempenho quando o modelo é avaliado em dados provenientes de um domínio diferente daquele utilizado no treinamento. Em ambos os cenários, a acurácia permaneceu próxima de 0.50, indicando desempenho semelhante a uma classificação aleatória balanceada.

Uma análise mais detalhada dos resultados revela que o modelo apresentou forte tendência em favorecer a classe *Fake*. No experimento FakeBR $\rightarrow$ LupaAFP, o recall da classe *Fake* atingiu valor próximo de 1.00, enquanto a classe *True* apresentou desempenho praticamente nulo.

Esse comportamento sugere que o modelo aprendeu padrões específicos presentes na base de treinamento, mas não conseguiu generalizar adequadamente para distribuições textuais distintas. Esse fenômeno caracteriza um problema de *dataset shift*, no qual diferenças estruturais, semânticas e estilísticas entre os conjuntos de dados impactam diretamente o desempenho do classificador.

Além disso, o experimento LupaAFP $\rightarrow$ FakeBR apresentou comportamento semelhante, reforçando a hipótese de que o modelo desenvolveu forte dependência das características particulares dos datasets utilizados durante o treinamento.

C. Comparação com Baseline MLP

Com o objetivo de investigar se o problema observado nos experimentos cross-domain estava relacionado especificamente à arquitetura do BERTimbau, foram realizados experimentos adicionais utilizando um classificador MLP com representação TF-IDF.

A Tabela IV apresenta os resultados intra-domain obtidos com o MLP.

Tabela IV
RESULTADOS INTRA-DOMAIN UTILIZANDO MLP

Dataset	Acurácia	F1 Macro
FakeBR	0.95	0.95
LupaAFP	0.67	0.67
FactckBR	0.52	0.52

Na base FakeBR, o MLP apresentou desempenho superior ao BERTimbau em cenário intra-domain. Entretanto, nas bases LupaAFP e FactckBR, o BERTimbau apresentou resultados superiores. Esse comportamento sugere que a abordagem TF-IDF pode capturar padrões lexicais altamente específicos em bases mais homogêneas, enquanto o BERTimbau apresenta maior capacidade de representação contextual em cenários mais variados.

A Tabela V apresenta os resultados cross-domain obtidos com o MLP.

Tabela V
RESULTADOS CROSS-DOMAIN UTILIZANDO MLP

Treino → Teste	Acurácia	F1 Macro
FakeBR → LupaAFP	0.50	0.35
LupaAFP → FakeBR	0.52	0.48

Os resultados obtidos com o MLP apresentaram comportamento semelhante ao observado nos experimentos utilizando o BERTimbau. Embora tenham ocorrido pequenas variações, o desempenho geral permaneceu limitado em tarefas de generalização entre domínios. Esse resultado reforça a hipótese de que o principal problema identificado não está exclusivamente relacionado à arquitetura do modelo, mas às diferenças existentes entre as distribuições textuais das bases utilizadas.

D. Experimentos Multi-Domínio

Com o objetivo de investigar o impacto da diversidade de dados na capacidade de generalização do modelo, foram realizados experimentos utilizando combinações de múltiplas bases de dados durante o treinamento.

Nos experimentos multi-domínio, duas bases foram combinadas para treinamento e a terceira foi utilizada integralmente como conjunto de teste, permitindo avaliar se a diversidade de domínios no treinamento contribui para melhorar a generalização.

No experimento Super Mix, os dados provenientes das três bases foram unificados e posteriormente divididos em 80% para treino e 20% para teste, utilizando divisão estratificada para manter a proporção das classes Fake e True.

A Tabela VI apresenta os resultados obtidos utilizando o modelo *BERTimbau* em cenários de treinamento multi-domínio.

Tabela VI
RESULTADOS DE GENERALIZAÇÃO UTILIZANDO TREINAMENTO MULTI-DOMÍNIO

Treino → Teste	Acurácia	F1 Macro
FakeBR + LupaAFP → FactckBR	0.65	0.64
FakeBR + FactckBR → LupaAFP	0.64	0.64
LupaAFP + FactckBR → FakeBR	0.50	0.33

Os resultados demonstram que a utilização de múltiplos domínios durante o treinamento contribuiu para melhorar a capacidade de generalização do modelo em determinados cenários.

O experimento FakeBR + FactckBR → LupaAFP apresentou acurácia de 0.64, valor superior ao experimento cross-domain tradicional FakeBR → LupaAFP, que havia obtido acurácia próxima de 0.50. Resultado semelhante foi observado no cenário FakeBR + LupaAFP → FactckBR, no qual o modelo atingiu acurácia de 0.65.

Entretanto, o experimento LupaAFP + FactckBR → FakeBR continuou apresentando desempenho limitado, com acurácia próxima de 0.50. Esse comportamento indica que determinadas características presentes na base FakeBR podem diferir significativamente das demais bases utilizadas, dificultando a transferência de aprendizado entre domínios.

Os experimentos reforçam a hipótese de que o fenômeno de *dataset shift* exerce forte influência sobre modelos de classificação de fake news. Além disso, os resultados indicam que estratégias de treinamento multi-domínio podem representar uma alternativa promissora para aumentar a robustez de modelos de linguagem em cenários reais.

E. Experimento Super Mix

Com o objetivo de avaliar o efeito da combinação completa das bases, foi realizado um experimento utilizando as três bases de dados em um único conjunto de treinamento, denominado *Super Mix*.

Nesse cenário, os dados provenientes das bases FakeBR, LupaAFP e FactckBR foram unificados e posteriormente divididos em conjuntos de treino e teste utilizando divisão estratificada.

A Tabela VII apresenta os resultados obtidos no experimento Super Mix.

Tabela VII
RESULTADOS DO EXPERIMENTO SUPER MIX UTILIZANDO BERTIMBAU

Classe	Precisão	Recall	F1-score
Fake	0.91	0.85	0.87
True	0.86	0.91	0.88

O modelo atingiu acurácia global de 0.88, apresentando desempenho significativamente superior aos experimentos cross-domain e comparável aos melhores resultados intra-domain.

A matriz de confusão obtida é apresentada na Equação 1.

$$\begin{bmatrix} 1091 & 198 \\ 114 & 1175 \end{bmatrix} \tag{1}$$

Os resultados sugerem que a utilização de múltiplos domínios durante o treinamento contribui para aumentar a robustez do modelo, permitindo maior exposição a diferentes padrões linguísticos, estruturas textuais e estilos de escrita.

Diferentemente dos experimentos cross-domain, nos quais o modelo apresentou forte dependência das características específicas da base de treinamento, o experimento Super Mix demonstrou maior equilíbrio entre as classes e melhor capacidade de adaptação a diferentes distribuições de dados.

F. Discussão sobre Diferenças entre os Datasets

A caracterização estatística dos datasets revelou diferenças expressivas entre as bases utilizadas. O FakeBR apresentou textos substancialmente maiores, com média superior a 360 palavras por notícia, enquanto LupaAFP e FactckBR apresentaram textos mais curtos, com médias próximas de 17 e 22 palavras, respectivamente.

Essa diferença estrutural ajuda a explicar parte da dificuldade observada nos experimentos cross-domain. Um modelo treinado em textos longos, com maior contexto e vocabulário mais amplo, pode não aprender padrões transferíveis para textos curtos e verificativos. Da mesma forma, um modelo treinado em textos curtos pode não capturar adequadamente estruturas narrativas mais longas.

Além do tamanho dos textos, também foram observadas diferenças temáticas entre as bases. Termos associados à política nacional e operações como Lava Jato apareceram com maior frequência em bases derivadas do FakeBR, enquanto temas relacionados a COVID-19, redes sociais e checagem de conteúdo apareceram com maior presença em bases de fact-checking. Isso sugere que os modelos podem estar aprendendo não apenas características associadas à veracidade, mas também padrões temáticos e temporais específicos de cada base.

Portanto, os resultados indicam que altos valores de acurácia em cenários intra-domain devem ser interpretados com cautela. Embora indiquem boa adaptação ao conjunto de dados utilizado, não garantem necessariamente robustez em cenários reais, nos quais as notícias podem apresentar diferentes temas, estilos e estruturas textuais.

G. Análise Qualitativa dos Datasets e Resultados

Além da análise quantitativa, foi realizada uma etapa de auditoria qualitativa utilizando o GPT-4o mini como ferramenta auxiliar de interpretação das decisões do classificador.

Os experimentos analisaram casos de acerto e erro do BERTimbau, buscando identificar fatores linguísticos e contextuais associados ao comportamento do modelo. As análises indicaram que, em diversos casos, o classificador apresentou forte dependência de padrões lexicais específicos presentes nas bases de treinamento.

Também foram observados conflitos entre as classificações produzidas pelo BERTimbau e as interpretações geradas pelo GPT-4o mini. Em determinados exemplos, o modelo de linguagem identificou ambiguidades contextuais e características semânticas que não foram adequadamente capturadas pelo classificador.

A Tabela VIII apresenta exemplos de aspectos extraídos pela etapa qualitativa.

Tabela VIII
EXEMPLOS DE ASPECTOS EXTRAÍDOS NA ANÁLISE QUALITATIVA

Base	**Aspectos identificados**
FakeBR	Linguagem política; narrativa longa; termos recorrentes
LupaAFP	Checagem direta; texto curto; contexto verificativo
FactckBR	Baixa amostragem; conteúdo factual; variação contextual

Os resultados qualitativos reforçam os achados observados nos experimentos quantitativos, indicando que o modelo apresenta limitações relevantes de generalização quando exposto a distribuições textuais distintas daquelas utilizadas durante o treinamento.

Dessa forma, a utilização de modelos de linguagem de grande porte como ferramenta auxiliar de interpretabilidade mostrou-se útil para ampliar a compreensão do comportamento do classificador, especialmente em cenários de erro e conflito semântico.

V. CONCLUSÃO

Este trabalho investigou o uso de modelos de linguagem para classificação automática de fake news em português, utilizando diferentes bases de dados com características textuais distintas.

Os resultados obtidos demonstraram que o modelo BERTimbau apresentou desempenho elevado em cenários intra-domain, atingindo acurácia de 0.93 na base FakeBR, 0.80 na base LupaAFP e 0.65 na base FactckBR. Esses resultados indicam que o modelo foi capaz de aprender padrões específicos dos conjuntos de dados utilizados durante o treinamento.

Entretanto, os experimentos cross-domain evidenciaram forte limitação de generalização. Nos cenários em que o treinamento e a avaliação ocorreram em domínios diferentes, o desempenho caiu para valores próximos de 0.50 de acurácia, comportamento semelhante ao observado em classificações aleatórias balanceadas.

Os experimentos também demonstraram que o problema não está relacionado exclusivamente à arquitetura utilizada. Os testes realizados com o classificador MLP apresentaram comportamento semelhante ao BERTimbau em cenários de mudança de domínio, reforçando a hipótese de que o principal fator associado à queda de desempenho está relacionado ao fenômeno de *dataset shift*.

As análises estatísticas realizadas sobre os datasets revelaram diferenças significativas entre as bases utilizadas, incluindo tamanho médio das notícias, diversidade de vocabulário e temas predominantes. Enquanto o FakeBR apresentou textos longos e vocabulário amplo, as bases LupaAFP e FactckBR apresentaram estruturas textuais mais curtas e objetivas. Essas diferenças contribuíram para explicar parcialmente as dificuldades de generalização observadas nos experimentos cross-domain.

Além disso, os experimentos multi-domínio demonstraram que a combinação de diferentes bases durante o treinamento contribui para melhorar a robustez do modelo, permitindo maior exposição a diferentes padrões linguísticos e contextuais.

A utilização do GPT-4o mini como ferramenta auxiliar de análise qualitativa permitiu examinar padrões textuais, ambiguidades semânticas e diferenças entre os datasets, contribuindo para interpretar os resultados experimentais. Essa etapa, entretanto, não foi tratada como explicação formal dos mecanismos internos do modelo, mas como uma análise complementar das características linguísticas e contextuais presentes nos dados..

A. Trabalhos Futuros

Como trabalhos futuros, sugere-se a investigação de técnicas de adaptação de domínio (*domain adaptation*) aplicadas à detecção de fake news em português, buscando reduzir os impactos causados pelo *dataset shift* observado nos experimentos.

Também podem ser exploradas estratégias de aumento de dados (*data augmentation*), utilização de modelos maiores baseados em transformadores e métodos de ajuste fino mais eficientes, como LoRA (*Low-Rank Adaptation*).

Outra possibilidade consiste na utilização de arquiteturas híbridas combinando recuperação de informação e modelos de linguagem, como abordagens baseadas em *Retrieval-Augmented Generation* (RAG), permitindo incorporar contexto externo durante a classificação.

Além disso, futuras pesquisas podem complementar a análise qualitativa com técnicas formais de Inteligência Artificial Explicável, como SHAP, LIME, análise de atenção, explicações token a token e comparação entre diferentes modelos de linguagem. Essas abordagens permitiriam investigar de forma mais direta quais elementos textuais influenciam as decisões dos classificadores.

Por fim, recomenda-se a construção de novos datasets em português contendo maior diversidade temática, temporal e estrutural, permitindo avaliações mais robustas da capacidade de generalização de modelos de detecção de fake news.

REFERÊNCIAS

[1] F. Souza, R. Nogueira, and R. Lotufo, "Bertimbau: Pretrained bert models for brazilian portuguese," in *Intelligent Systems*. Springer, 2020, pp. 403–417.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[3] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[4] R. A. Monteiro, R. L. S. Santos, T. A. S. Pardo, T. A. de Almeida, E. E. S. Ruiz, and O. A. Vale, "Contributions to the study of fake news in portuguese: New corpus and automatic detection results," in *Computational Processing of the Portuguese Language*. Springer, 2018, pp. 324–334.

[5] J. Moreno and et al., "Factck.br: A new dataset to study fake news," in *Proceedings of the Brazilian Symposium on Multimedia and the Web*, 2019.

[6] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[7] R. A. Monteiro, R. L. S. Santos, T. A. S. Pardo, T. A. de Almeida, E. E. S. Ruiz, and O. A. Vale, "Contributions to the study of fake news in portuguese: New corpus and automatic detection results," in *Computational Processing of the Portuguese Language*. Springer International Publishing, 2018, pp. 324–334.