Detecção de Fake News em Português: Análise Comparativa entre Métodos de Representação em Português, Inglês e Multilíngues

Camila B. Vieira[1], **José Vinicius de S. Souza**[1], **George D. C. Cavalcanti**[1]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Recife – PE – Brasil

{jvss2,cbv2,gdcc}@cin.ufpe.br

Abstract. *This study investigates the effectiveness of different text representation methods in detecting fake news in Portuguese. We evaluated models trained in Portuguese, English, and multilingual settings, using MLP, RFC, and SVC classifiers on the balanced FAKE.BR CORPUS dataset. We also analyzed instance hardness to measure the classification difficulty of instances. The results show that, among the Portuguese models, only BERTimbau achieved significant performance. Despite the higher computational cost, multilingual models demonstrated advantages, while those trained in English performed competitively but remained inferior. The source code is available in the repository* `https://github.com/camilab-vieira/compare-embeddings.git`.

Resumo. *Este estudo investiga a eficácia de diferentes métodos de representação textual na detecção de fake news em português. Foram avaliados modelos treinados em português, inglês e multilíngues, utilizando classificadores MLP, RFC e SVC no conjunto balanceado FAKE.BR CORPUS. Também analisamos a instance hardness para mensurar a dificuldade de classificação das instâncias. Os resultados mostram que, entre os modelos em português, apenas o BERTimbau apresentou desempenho expressivo. Apesar do maior custo computacional, os modelos multilíngues revelaram vantagens, enquanto os treinados em inglês tiveram desempenho competitivo, mas inferior. O código-fonte está disponível no repositório* `https://github.com/camilab-vieira/compare-embeddings.git`.

1. Introdução

A desinformação é apontada como o maior risco global para os próximos dois anos [World Economic Forum 2024], devido à sua capacidade de influenciar processos eleitorais, intensificar a polarização política e ameaçar a estabilidade social. O combate à desinformação, no entanto, representa um desafio complexo: por um lado, medidas excessivas podem resultar em censura e restrições à liberdade de expressão; por outro, a falta de ação favorece a propagação de informações falsas, ampliando ainda mais os riscos associados à sua disseminação.

Diante desse cenário, cresce a necessidade de soluções automatizadas para auxiliar os *fact-checkers*—especialistas responsáveis por verificar a veracidade das informações. Esses profissionais, geralmente vinculados a grandes veículos de comunicação, enfrentam

um volume crescente de notícias falsas, tornando inviável a verificação manual de todo o conteúdo compartilhado diariamente [Eiseler 2019].

O desenvolvimento dessas soluções, porém, apresenta desafios significativos [Diego N. E. Silva 2019], sobretudo para idiomas com menos recursos computacionais e conjuntos de dados disponíveis para treinamento, como o português [Almeida et al. 2024]. A escassez de dados rotulados e a menor disponibilidade de modelos especializados dificultam a criação de abordagens eficazes. Em contraste, o inglês dispõe de um ecossistema de pesquisa mais robusto, com vastos volumes de dados para treinamento e modelos continuamente aprimorados, favorecendo o desenvolvimento de soluções altamente otimizadas para tarefas como a detecção de *fake news*.

Para mitigar essas limitações, diferentes estratégias vêm sendo exploradas. Alguns modelos de representação textual foram ajustados especificamente para o português, como BERTimbau, TeenyTinyLlama e Tucano [Souza et al. 2020, Corrêa et al. 2024a, Corrêa et al. 2024b]. Paralelamente, modelos pré-treinados para múltiplos idiomas, como mBART, mBERT e XLM-RoBERTa [Devlin et al. 2019, Conneau et al. 2020, Liu and Lapata 2020], são investigados como alternativas para aproveitar padrões compartilhados entre línguas. Nesse contexto, torna-se essencial avaliar se os modelos adaptados ao português são capazes de alcançar um desempenho competitivo em comparação com as abordagens multilíngues ou com os modelos originalmente desenvolvidos para o inglês.

Neste estudo, realizamos experimentos utilizando diferentes métodos de representação textual e classificadores. Para cada um dos *embeddings* testados, aplicamos os classificadores *Multi-Layer Perceptron* (MLP), *Random Forest Classifier* (RFC) e *Support Vector Classifier* (SVC) no conjunto de dados balanceado FAKE.BR CORPUS [Santos et al. 2018], que contém 7.200 notícias de tópicos variados publicadas até 2018, cada uma com pelo menos 100 palavras. O código-fonte e os resultados complementares deste trabalho estão disponíveis no repositório `https://github.com/camilab-vieira/compare-embeddings.git`.

As principais contribuições deste estudo são:

- Avaliação de diferentes métodos de representação pré-treinados para a detecção de *fake news* em português brasileiro;
- Análise da combinação entre métodos de representação e classificadores;
- Investigação da dificuldade para classificação das instâncias de acordo com a forma como são representadas, *instance hardness* [Smith 2009], considerando também o custo computacional do processo de extração desses *embeddings*.

2. Trabalhos Relacionados

A detecção automática de notícias falsas tem sido um tema amplamente estudado na área de aprendizado de máquina e processamento de linguagem natural. Diversos estudos exploram diferentes abordagens para identificar padrões linguísticos, estilísticos e estruturais que distinguem textos verídicos de falsos [Braz and Digiampietri 2024, Reis and Benevenuto 2022, Graciano Neto et al. 2024]. Um dos principais desafios dessa tarefa é garantir a generalização dos modelos diante da variabilidade textual e das estratégias utilizadas para a disseminação de desinformação.

Entre os trabalhos que abordam essa temática, destaca-se o estudo de [Sousa et al. 2022]. Os autores utilizam o FAKE.BR CORPUS para treinar uma abordagem híbrida que combina Redes Neurais Convolucionais para extração de características semânticas e algoritmos tradicionais de aprendizado de máquina para análise de metadados. Os resultados obtidos demonstram uma acurácia de 97%, evidenciando a eficácia da fusão de técnicas baseadas em aprendizado profundo e métodos clássicos para a classificação de notícias falsas em português. No entanto, este estudo tem uma alta complexidade computacional associada a sua abordagem híbrida, que pode dificultar sua implementação em cenários de tempo real ou com recursos computacionais limitados.

O desempenho de modelos BERT foram avaliados na classificação de notícias falsas em português, utilizando o dataset FAKE.BR CORPUS e outros [Pires and e Silva 2024]. Seus resultados indicaram que o BERT atingiu uma acurácia de 0,897, o mBERT alcançou 0,938 e o BERTimbau obteve a melhor performance com 0,950. O estudo reforça a importância de modelos pré-treinados na língua-alvo. Pretendemos expandir a análise para além da família BERT, comparando o desempenho de diferentes *embeddings* na tarefa de detecção de notícias falsas.

Outro estudo relevante é o de [Vicentini 2023]. Neste trabalho, a autora analisa diferentes métodos de interpretabilidade aplicados a modelos de linguagem para a detecção de *fake news*. A pesquisa explora como técnicas como *LIME* e *SHAP* podem contribuir para entender as decisões tomadas por modelos baseados em aprendizado profundo, promovendo maior transparência e confiabilidade na classificação de notícias falsas. No entanto, o estudo se concentra apenas em algumas técnicas de explicabilidade, sem explorar outras abordagens emergentes, como a análise do *Instance Hardness*, que poderiam oferecer insights adicionais sobre o comportamento dos modelos.

3. Metodologia

A metodologia deste estudo avalia a eficácia de diferentes representações textuais e algoritmos de classificação na detecção de *fake news* em português brasileiro, utilizando o FAKE.BR CORPUS. São descritas a composição do *dataset*, as representações textuais adotadas - monolíngues (português e inglês) e multilíngues -, os classificadores utilizados e os ajustes de hiperparâmetros. Por fim, são apresentadas as métricas de avaliação: acurácia, *F1-Score* e tempo de execução.

3.1. Dataset

O FAKE.BR CORPUS [Santos et al. 2018] foi desenvolvido para a análise de notícias verdadeiras e falsas em português brasileiro. Ele contém 7.200 notícias, sendo 3.600 verdadeiras e 3.600 falsas. Parte das notícias falsas foi manualmente elaborada para espelhar a estrutura e o estilo das verdadeiras, garantindo um alinhamento fiel com suas contrapartes, mas com informações distorcidas ou inventadas. O conjunto abrange diversos tópicos e inclui apenas textos com mais de 100 palavras, publicados até 2018.

Os dados já estão pré-processados e estratificados em conjuntos de treinamento e teste (80/20), facilitando a aplicação de modelos de aprendizado de máquina. O balanceamento das classes é um aspecto crucial, pois evita viés para uma das categorias e permite que os modelos aprendam a distinguir com maior precisão entre notícias verdadeiras e falsas.

3.2. Métodos de Representação

A escolha dos métodos de representação influencia diretamente a qualidade dos *embeddings* gerados e, consequentemente, o desempenho dos classificadores [Farhangian et al. 2024]. Em todos os métodos utilizados, empregamos a versão base ou menor dos modelos, que possui 768 dimensões. A exceção é o mBART, que utiliza 1024 dimensões. Além disso, todos os textos processados, mesmo para os modelos monolíngues em inglês, são exclusivamente em português

- **Modelos em português:** São ajustados especificamente para o idioma por meio de *fine-tuning*, permitindo representações mais adequadas para tarefas em português, embora dependam de dados de treinamento limitados.
  - **BERTimbau:** é uma adaptação do BERT para o português, treinado em um grande corpus do idioma. Sua arquitetura bidirecional permite a captação de relações contextuais profundas, tornando-o uma forte referência para tarefas de NLP (*Natural Language Processing*) em português. Seu treinamento monolíngue favorece representações mais precisas para textos neste idioma [Souza et al. 2020].
  - **TeenyTinyLlama:** é uma versão compacta do Llama 2, projetada com restrições computacionais e orçamentárias. Além do seu tamanho reduzido, essas limitações podem afetar a expressividade dos *embeddings*, especialmente em tarefas que exigem alta discriminação semântica [Corrêa et al. 2024a].
  - **Tucano:** é um LLM (*Large Language Model*) otimizado para eficiência computacional sem comprometer a qualidade dos *embeddings*. Seu design busca equilibrar custo e desempenho, sendo uma alternativa viável para aplicações que exigem modelos leves sem perda significativa de representatividade [Corrêa et al. 2024b].
- **Modelos em inglês:** Aproveitam grandes quantidades de dados e avanços recentes no campo de NLP, mas podem não capturar plenamente as particularidades do português.
  - **BERT:** é um modelo bidirecional pré-treinado com *Masked Language Modeling* (MLM) e *Next Sentence Prediction* (NSP), permitindo representações contextuais ricas. Sua robustez o torna uma base sólida para diversas tarefas de NLP [Devlin et al. 2019].
  - **RoBERTa:** aprimora o BERT ao remover a tarefa de NSP e ao empregar um treinamento mais extenso com mascaramento dinâmico. Como resultado, gera *embeddings* mais precisos e robustos, o que pode beneficiar modelos que requerem maior expressividade semântica [Liu et al. 2019].
  - **BART:** combina características do BERT e do GPT ao utilizar um modelo *seq2seq* treinado para reconstrução de texto corrompido. Sua arquitetura é especialmente útil para tarefas como *paraphrasing*, geração de texto e adaptação de *embeddings* a diferentes domínios [Lewis et al. 2019].
- **Modelos multilíngues:** São treinados em múltiplos idiomas, incorporando estruturas interlinguísticas que podem beneficiar línguas menos representadas. No entanto, esse benefício vem ao custo de maior complexidade e exigência computacional.

- **mBERT:** é uma versão do BERT treinada em múltiplos idiomas, permitindo transferências linguísticas úteis. Embora tenha sido projetado para funcionar em diversos contextos, sua capacidade de representação pode ser diluída quando comparada a modelos monolíngues especializados [Devlin et al. 2019].
- **XLM-RoBERTa:** expande o modelo BERT ao incorporar um pré-treinamento robusto em múltiplas línguas, aumentando sua capacidade de generalização. No entanto, esse aprimoramento exige maior poder computacional, o que pode impactar a viabilidade de uso em determinados cenários [Conneau et al. 2020].
- **mBART:** segue um paradigma de reconstrução textual, sendo particularmente útil em tarefas que exigem transferência entre idiomas. Sua estrutura *seq2seq* e sua abordagem de *denoising* o tornam eficaz para adaptação linguística, o que pode beneficiar representações em português mesmo sem um treinamento monolíngue dedicado [Liu and Lapata 2020].

3.3. Classificadores

Nesta seção, definimos os classificadores utilizados e seus hiperparâmetros. Os métodos selecionados incluem modelos baseados em árvores de decisão, redes neurais artificiais e vetores de suporte, garantindo diversidade na modelagem das relações presentes nos dados. Optou-se por não empregar arquiteturas profundas devido à sua elevada complexidade computacional, priorizando abordagens mais leves e eficientes, compatíveis com os recursos disponíveis. Os experimentos foram conduzidos utilizando a biblioteca *scikit-learn* [Pedregosa et al. 2011].

A escolha adequada dos hiperparâmetros é essencial para otimizar o desempenho dos classificadores [Bahmani et al. 2025]. Para isso, utilizamos no conjunto de treino, o *GridSearchCV*, uma técnica que realiza uma busca exaustiva sobre combinações pré-definidas de hiperparâmetros, avaliando o desempenho por validação cruzada. O critério de seleção foi a métrica *F1-Score*, garantindo um equilíbrio entre precisão e *recall*. A Tabela 1 apresenta os hiperparâmetros testados para cada modelo: MLP, RFC e SVC.

3.4. Métricas

Para avaliar o desempenho dos modelos na detecção de *fake news*, utilizamos acurácia e *F1-Score* como principais métricas. A acurácia, calculada como a proporção de previsões corretas em relação ao total de amostras, é adequada para este estudo, pois o conjunto de dados é balanceado, garantindo que a métrica não seja enviesada por distribuições desiguais entre classes. O *F1-Score*, que representa a média harmônica entre precisão e *recall*, complementa a análise ao fornecer uma visão mais detalhada sobre os erros do modelo. Além disso, avaliamos o tempo de execução, considerando tanto o tempo necessário para gerar as representações dos textos (extração dos *embeddings*), quanto o tempo de treinamento dos classificadores, fatores essenciais para determinar a viabilidade prática das abordagens testadas.

4. Experimentos

Nesta seção, apresentamos os resultados dos experimentos realizados com diferentes métodos de representação e classificadores. Os resultados resumidos em tabelas uti-

Tabela 1. Hiperparâmetros testados para os classificadores MLP, RFC e SVC utilizando GridSearchCV com validação cruzada.

Modelo	Hiperparâmetros
MLP	*activation*: [relu, logistic], *solver*: [adam, lbfgs]
RFC	*bootstrap*: [True, False], *max_depth*: [5, 10, 20, 30], *max_features*: [auto, sqrt, log2], *min_samples_leaf*: [1, 2, 4], *min_samples_split*: [2, 5, 10], *n_estimators*: [50, 100, 200], *criterion*: [gini, entropy]
SVC	*kernel*: [rbf], *gamma*: [1, 0.1, 0.01, 0.001, 0.0001], *C*: [0.1, 1, 10, 100, 1000]

lizam uma codificação por cores para facilitar a análise: o vermelho indica o melhor desempenho, o azul o segundo melhor, e o violeta o terceiro. Os experimentos foram realizados utilizando um nó do Cluster Apuana, pertencente ao Centro de Informática da Universidade Federal de Pernambuco (UFPE). Mais informações sobre o ambiente de execução, incluindo detalhes técnicos da infraestrutura, estão disponíveis em `https://apuana.cin.ufpe.br`.

Na Tabela 2, podemos visualizar apenas o tempo para extrair *embeddings*. Observa-se que mBART requer o maior tempo, o que pode limitar seu uso em cenários com restrições computacionais, enquanto BERTimbau e BART são os mais rápidos. TeenyTinyLlama e Tucano apresentam tempos elevados, apesar de suas arquiteturas menores, enquanto variantes multilíngues, como mBERT e XLM-RoBERTa, situam-se em um intervalo intermediário. Essas diferenças refletem variações arquiteturais e de eficiência computacional, ressaltando a importância de considerar o custo computacional na escolha do modelo.

Tabela 2. Tempo de processamento (em horas) para extração dos *embeddings* ao gerar representações para os textos.

Embedding	**Tempo**
BERTIMBAU	1,43
TEENNYTINYLLAMA	5,55
TUCANO	6,53
BERT	2,29
ROBERTA	2,29
BART	1,72
MBERT	3,08
MBART	10,38
XLM-ROBERTA	2,62

4.1. Análise Comparativa

Os experimentos, apresentados na Tabela 3, evidenciam diferenças significativas no desempenho entre os métodos de representação e os classificadores avaliados. O tempo reportado corresponde à soma do tempo de geração dos *embeddings* e do treinamento dos respectivos modelos.

O BERTimbau, consolidado na língua portuguesa, obteve bons resultados, superando os *embeddings* ajustados recentemente para o português, como TeenyTinyLlama e Tucano, que apresentaram desempenho insatisfatório em todos os classificadores. Esses dois últimos, apesar do *fine-tuning*, mostraram dificuldades na representação textual, além de demandarem maior tempo de processamento. Isso pode estar relacionado à menor quantidade e qualidade dos dados utilizados no treinamento, à limitação arquitetural dos modelos ou a um *fine-tuning* insuficiente para capturar padrões relevantes para a detecção de *fake news*.

Entre os classificadores, o SVC apresentou o melhor desempenho, enquanto o RFC teve os piores tempos de treinamento, tornando-o menos eficiente. Essa tendência foi observada na maioria dos métodos de representação, possivelmente influenciada pelos hiperparâmetros selecionados no *GridSearch*. Tanto o tempo de treinamento quanto a solução ótima podem ter sido afetados pelas configurações testadas.

Os modelos pré-treinados em inglês mostraram resultados competitivos, indicando que avanços disponíveis apenas nessa língua podem ser aproveitados para a detecção de *fake news* em português. No entanto, seus desempenhos foram inferiores aos dos modelos ajustados para o português ou múltiplas línguas, reforçando a importância do *fine-tuning*.

Os modelos multilíngues, como mBERT, XLM-RoBERTa e mBART, obtiveram desempenhos próximos ao do BERTimbau, mas com um custo computacional maior. O mBART, em particular, superou todos os demais *embeddings*, inclusive o BERTimbau, mas com um tempo de treinamento muito superior, tornando-se uma opção mais custosa. Isso pode ser atribuído, em parte, ao fato de o mBART utilizar 1024 dimensões, enquanto os outros modelos, como o BERTimbau, utilizam 768 dimensões.

Os resultados indicam que o *fine-tuning* em português melhora significativamente o desempenho. Modelos multilíngues são uma alternativa viável e podem superar representações monolíngues, mas exigem maior capacidade computacional. Entre os classificadores, o SVC se destaca como a melhor escolha ao considerar tanto métricas de desempenho quanto custos computacionais.

A Figura 1 evidencia que, embora o mBART apresente alta acurácia, seu tempo de extração é significativamente elevado, tornando-o uma opção menos viável em cenários com restrições computacionais. Observa-se também que Tucano e TeenyTinyLlama destoam dos demais modelos, combinando tempos elevados com acurácias inferiores, especialmente em comparação com BERTimbau. Ao desconsiderar esses *outliers*, percebe-se uma tendência de que os modelos multilíngues alcançam acurácias mais altas, mas a um custo computacional maior. No entando, o RoBERTa requer mais tempo sem apresentar ganhos de acurácia em relação aos demais modelos multilíngues.

Os resultados obtidos com BERTimbau e SVC (97,22%) e mBART e SVC

Tabela 3. Resultados considerando diferentes métodos de representação e classificadores. São apresentadas a acurácia, o *F1-Score* e o tempo de inferência (em horas) para cada configuração testada.

Português

Embedding	Modelo	Acurácia	F1 score	Tempo
BERTIMBAU	MLP	96,88	96,87	1,45
BERTIMBAU	RFC	93,68	93,68	1,76
BERTIMBAU	SVC	97,22	97,22	1,45
TEENNYTINYLLAMA	MLP	49,79	33,24	5,55
TEENNYTINYLLAMA	RFC	49,79	33,24	5,88
TEENNYTINYLLAMA	SVC	49,79	33,24	5,59
TUCANO	MLP	66,88	66,76	6,58
TUCANO	RFC	66,67	66,22	6,89
TUCANO	SVC	67,64	67,53	6,59

Inglês

Embedding	Modelo	Acurácia	F1 score	Tempo
BERT	MLP	91,60	91,58	2,34
BERT	RFC	90,49	90,47	2,55
BERT	SVC	92,99	92,98	2,33
ROBERTA	MLP	95,21	95,21	3,55
ROBERTA	RFC	90,28	90,27	3,76
ROBERTA	SVC	95,28	95,28	3,54
BART	MLP	87,99	87,98	1,78
BART	RFC	85,63	85,61	2,00
BART	SVC	88,75	88,75	1,78

Multilíngue

Embedding	Modelo	Acurácia	F1 score	Tempo
MBERT	MLP	95,90	95,90	3,12
MBERT	RFC	93,19	93,19	3,51
MBERT	SVC	96,39	96,39	3,12
XLM-ROBERTA	MLP	96,94	96,94	2,66
XLM-ROBERTA	RFC	93,06	93,05	3,03
XLM-ROBERTA	SVC	96,67	96,67	2,66
MBART	MLP	97,01	97,01	10,44
MBART	RFC	93,96	93,96	10,88
MBART	SVC	97,43	97,43	10,44

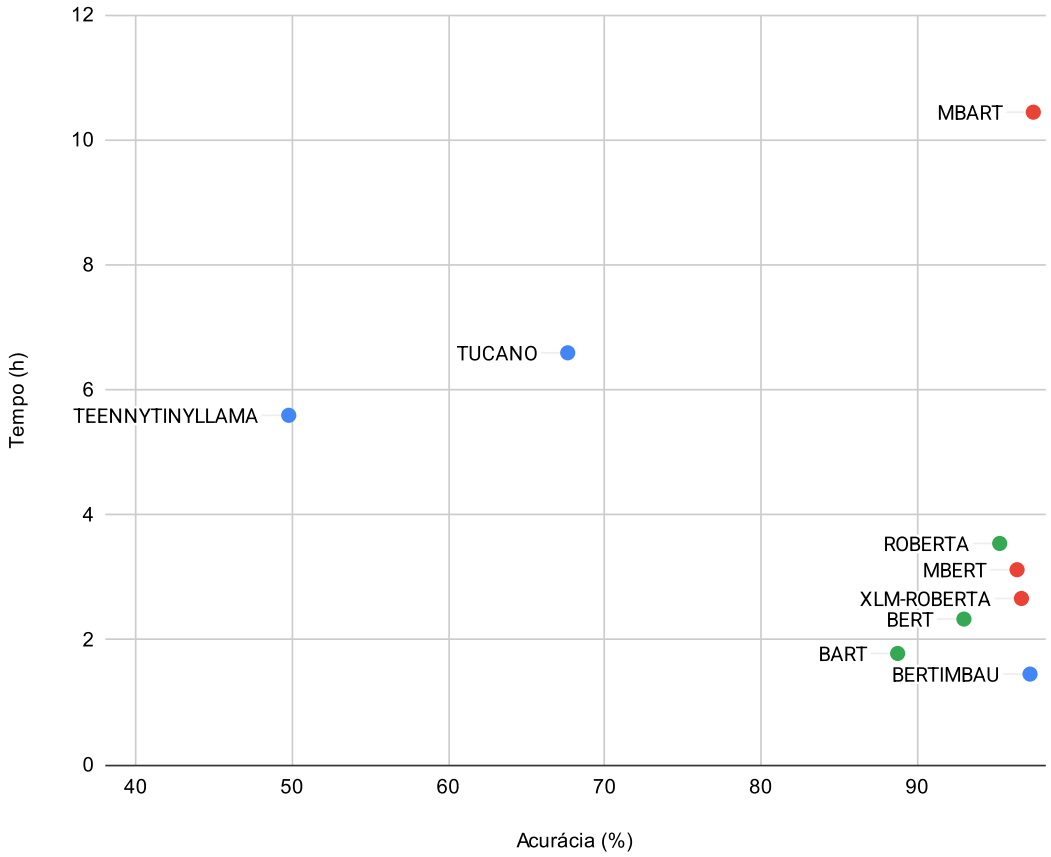

Figura 1. Acurácia por Tempo de Extração (em horas) com SVC – Embeddings em Português (azul), Inglês (verde) e Multilíngues (vermelho)

(97,43%) demonstram desempenho competitivo em relação à literatura. Estudos anteriores [Sousa et al. 2022], utilizando *GloVe* [Pennington et al. 2014] como representação, reportam 96,5% de acurácia para CNN e 97,5% para SVC com metadados, enquanto a combinação desses modelos atingiu 97,7%. Notavelmente, nossos experimentos, sem o uso de metadados, alcançaram acurácias próximas às melhores configurações reportadas, destacando a eficácia das representações baseadas em modelos de linguagem contextualizados em comparação com GloVe.

Modelos BERT foram avaliados para a classificação de fake news em português, destacando a superioridade do BERTimbau (0,950) sobre o mBERT (0,938) e o BERT (0,897), reforçando a importância de modelos ajustados para a língua-alvo [Pires and e Silva 2024]. Nossos experimentos corroboram essa tendência, mas também sugerem que nem todo modelo treinado em português é ideal, nem todo modelo multilíngue é inferior, indicando que a escolha do método de representação pode depender de fatores como a arquitetura, o processo de fine-tuning e outros aspectos do treinamento.

4.2. Análise do *Instance Hardness*

Instance Hardness (IH) [Smith 2009] quantifica a dificuldade de classificar corretamente uma instância. Um dos principais métodos para calculá-la é o *k-Disagreeing Neighbors*

(KDN), que mede a proporção de vizinhos mais próximos de uma instância que pertencem a classes diferentes. Valores altos indicam instâncias difíceis de classificar, enquanto valores baixos sugerem instâncias mais fáceis.

A Figura 2 exibe a distribuição cumulativa dos valores de KDN para cada método de representação. O gráfico do TeenyTinyLlama apresenta um comportamento anômalo, com uma grande concentração de valores extremos, sugerindo que esse modelo não consegue gerar *embeddings* adequados. Já o Tucano, como esperado, apresenta valores de KDN mais altos, indicando maior dificuldade de classificação. Por outro lado, as representações que obtiveram os melhores resultados, como BERTimbau, mBART e mBERT, possuem distribuições mais concentradas em baixos valores de KDN, reforçando a relação entre baixa dificuldade e melhor desempenho, conforme mostrado na Tabela 3.

Dessa forma, a investigação de IH fornece uma análise prévia do desempenho esperado dos *embeddings*, permitindo selecionar aqueles mais promissores antes mesmo do treinamento completo, evitando custos computacionais desnecessários.

5. Conclusão e Trabalhos Futuros

Observamos que a melhor combinação em termos de desempenho foi mBART com o classificador SVC, porém, devido ao maior tempo de processamento desse modelo, a combinação BERTimbau e SVC se destaca como a melhor alternativa, oferecendo um bom equilíbrio entre precisão e eficiência. Embora seja possível utilizar recursos desenvolvidos para o inglês, a adaptação específica para o português brasileiro resultou em uma melhoria significativa na qualidade da detecção de *fake news*. Esse resultado destaca a importância de adaptar as representações linguísticas às características específicas da língua em questão.

Para trabalhos futuros, planeja-se realizar uma análise mais abrangente, explorando outros métodos de representação, uma maior variação de hiperparâmetros e diferentes conjuntos de dados. Um foco importante será entender como a complexidade dos datasets (taxa de desbalanceamento, variabilidade, tamanho das instâncias...) influenciam nos resultados. Essa análise permitirá um melhor entendimento de como as características dos dados impactam o desempenho dos modelos e ajudará a refinar as abordagens para o combate a desinformação.

Referências

Almeida, R., Campos, R., Jorge, A., and Nunes, S. (2024). Indexing portuguese nlp resources with pt-pump-up. In *International Conference on Computational Processing of Portuguese*.

Bahmani, M., El Shawi, R., Potikyan, N., and Sakr, S. (2025). To tune or not to tune? an approach for recommending important hyperparameters for classification and clustering algorithms. *Future Generation Computer Systems*, 163:107524.

Braz, R. R. and Digiampietri, L. A. (2024). Detecção de fake news em domínios cruzados: Uma revisão sistemática. In *Brazilian Workshop on Social Network Analysis and Mining*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual re-

presentation learning at scale. In *Annual Meeting of the Association for Computational Linguistics*.

Corrêa, N. K., Falk, S., Fatimah, S., Sen, A., and De Oliveira, N. (2024a). Teenytinyllama: Open-source tiny language models trained in brazilian portuguese. *Machine Learning with Applications*, 16:100558.

Corrêa, N. K., Sen, A., Falk, S., and Fatimah, S. (2024b). Tucano: Advancing neural text generation for portuguese. arXiv. License: CC BY-NC-SA 4.0.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Diego N. E. Silva (2019). *Automating the Fact-Checking Task: Challenges and Directions*. PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn.

Eiseler, V. (2019). Redações devem adotar fact checking automatizado, escuta o público do isoj. *LatAm Journalism Review by the Knight Center*.

Farhangian, F., Cruz, R. M. O., and Cavalcanti, G. D. C. (2024). Fake news detection: Taxonomy and comparative study. *Information Fusion*, 103:102140.

Graciano Neto, V. V., Barbosa, J. R., Lima, E. A. d., Carvalho, S. T. d., and Venzi, S. (2024). A blockchain-based and ai-endorsed mechanism to support social networks on fake news containment. In *Brazilian Workshop on Social Network Analysis and Mining*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv.

Liu, Y. and Lapata, M. (2020). mBART: Multilingual denoising pre-training for neural machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Pires, V. B. and e Silva, D. G. (2024). Portuguese fake news classification with bert models. In *Encontro Nacional de Inteligência Artificial e Computacional*.

Reis, J. C. S. and Benevenuto, F. (2022). Detecção automática de desinformação em diferentes cenários: Eleições nos estados unidos e no brasil. In *Brazilian Workshop on Social Network Analysis and Mining*.

Santos, R. L. S., Monteiro, R. A., and Pardo, T. A. S. (2018). The fake.br corpus - a corpus of fake news for brazilian portuguese. In *International Conference on Computational Linguistics*.

Smith, M. R. (2009). *An Empirical Study of Instance Hardness*. PhD thesis, Brigham Young University.

Sousa, F., Barbosa, A., Oliveira, C., and Braga, R. (2022). Detecção de fake news em língua portuguesa combinando redes neurais convolucionais e algoritmos de aprendizagem de máquina. In *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417.

Vicentini, J. (2023). Comparando técnicas de explicabilidade sobre modelos de linguagem: um estudo de caso na detecção de notícias falsas. *Universidade Estadual Paulista*.

World Economic Forum (2024). Global risks report 2024.

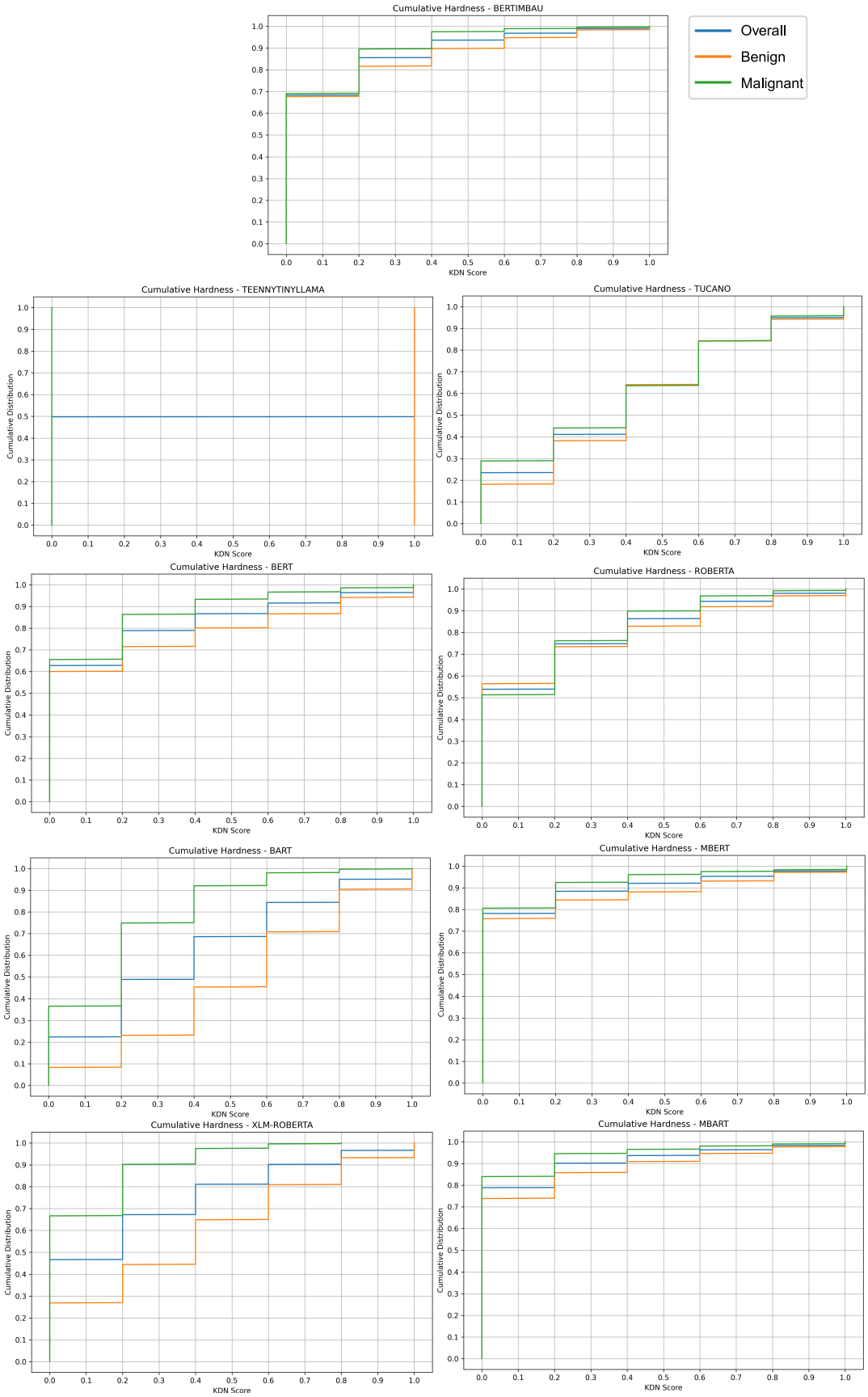

Figura 2. Distribuição cumulativa do *Instance Hardness* (IH), medido via KDN Score, para diferentes *embeddings*. Quanto maior o valor, mais difícil é classificar a instância.