Classificação de Fake News Utilizando Deep Learning: Uma Abordagem Multimodelos

1st Tenório, José Paulo Cauás
Centro de Informática
UFPE
Recife, Brasil
jpct@cin.ufpe.br

2nd Silva, Fagner Fernandes Candido da
Centro de Informática
UFPE
Recife, Brasil
ffcs@cin.ufpe.br

***Abstract*—The aim of this proposal is to apply Deep Learning models to classify news articles as either true or false. The approach involves using models based on different strategies such as Bag of Words, Embeddings, and BERT to extract textual features and perform binary classification of news content. The proposal will be validated using the ISOT Fake News Dataset, which contains 23,502 fake news articles and 21,417 real ones, with performance metrics including accuracy, precision, recall, and F1-score.**

***Index Terms*—deep learning, rnn, transformers, multimodel.**

I. Introdução

O advento da internet trouxe uma gama de impactos transformadores para a sociedade moderna, proporcionando acesso amplo e democratizado à informação. Esta revolução digital impulsionou empresas a digitalizarem seus produtos e serviços para se adaptarem às mudanças tecnológicas e manterem sua competitividade no mercado [2]. Como consequência desse avanço tecnológico, testemunhamos o surgimento massivo das redes sociais online (OSN - Online Social Networks), com destaque especial para as redes sociais multimídia (MSN - Multimedia Social Networks), que se concentram na experiência de compartilhamento de conteúdo multimídia [3].

Embora as redes sociais online e multimídia ofereçam vantagens significativas em termos de comunicação e avanço tecnológico, essas inovações também trouxeram impactos severos no aspecto social. A facilidade de criação e disseminação de conteúdo nas plataformas digitais criou um ambiente propício para a propagação de informações não verificadas, rumores e, consequentemente, fake news.

O artigo original [1] busca analisar o desempenho de diferentes abordagens de processamento de linguagem natural utilizando uma combinação entre diferentes modelos com word embeddings pré-treinados. Ele utilizou como base para essa análise um modelo híbrido de deep learning para detecção de fake news, combinando arquiteturas como CNN, Bidirectional LSTM e ResNet com word embeddings pré-treinados (Word2Vec, GloVe e fastText). Para mitigar desequilíbrios de classes, os autores aplicaram aumento de dados via back-translation (tradução reversa inglês-alemão-inglês). O pré-processamento incluiu etapas como remoção de pontuação, stopwords, lematização e tokenização.

Esse estudo, entretanto, revisita alguns modelos tradicionais como o Árvores Aleatórias (Random Forest), Naive Bayes, XGBoost, Regressão Logística com word embeddings específicos (Glove e Word2Vec) assim como grandes modelos de linguagem (Large Language Models - LLM), BERT com intuito de verificar o desempenho desses modelos na identificação de notícias falsas (fake news) como descrito na figura 1.

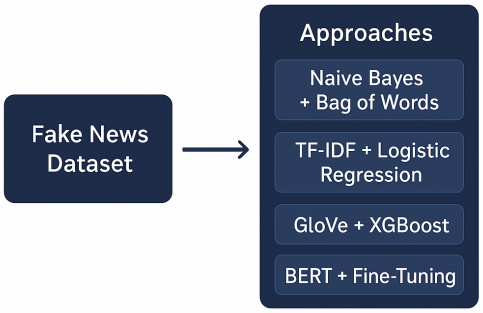

Fig. 1. Abordagem Multimodelos

Com relação ao artigo original, esse paper busca preencher alguns gaps de pesquisa relevantes como:

- explorar o optuna como ferramenta de tunning automatizado;
- incluir mecanismos de interpretabilidade através do LIME e SHAP;
- avaliar o desempenho de transformers no processo de identificação de fake news;

A. Caracterização do estudo

Os datasets utilizados para esse experimento foram quatro datasets públicos: ISOT Fake News, Fake News Dataset, Fake or Real News e Fake News Detection Dataset, totalizando 76.574 de registros analisados. A distribuição da quantidade de notícias em cada dataset está presente na figura 2.

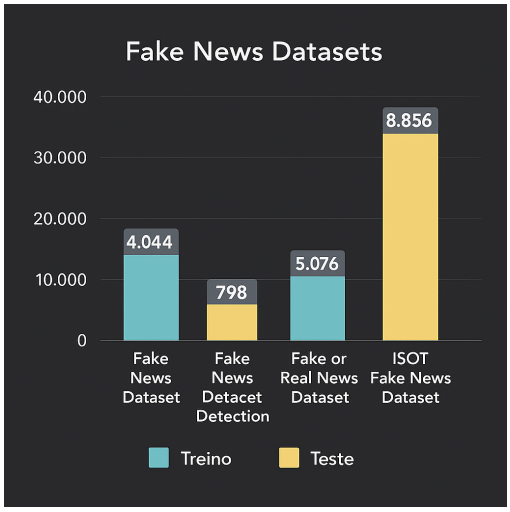

Fig. 2. Fake News Dataset

No artigo original, foi feito uma comparação com trabalhos anteriores que utilizaram TF-IDF, SVM, Random Forest e LSTMs unidirecionais. O desenho da arquitetura proposta está reprsentado na figura 3. Por sua vez, os resultados de Ahmed et al. [6] e Kaliyar et al. [7] foram replicados e comparados, atestando a superioridade do modelo proposto. Além disso, foram testadas combinações de embeddings (Word2Vec, GloVe, fastText) e arquiteturas (CNN, Bidirectional LSTM, ResNet), em conjunto com ajustes de hiperparâmetros (batch size: 32–512; otimizadores: Adamax destacou-se).

Nesse estudo foi utilizado um split de dados no padrão de 80% dos dados para treinamento e 20% para teste.

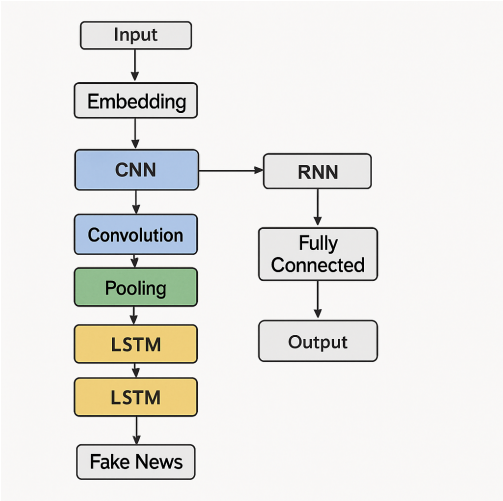

Fig. 3. Arquitetural LSTM Bidirecional

II. Reavaliando o estudo com uma proposta multimodelo

A crescente propagação de notícias falsas em plataformas digitais tem motivado o desenvolvimento de abordagens automatizadas de detecção baseadas em técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina. Este trabalho apresenta e compara cinco abordagens distintas para a classificação de notícias verdadeiras e falsas, utilizando diferentes formas de representação textual e algoritmos de aprendizado supervisionado. O objetivo é avaliar a eficácia de métodos clássicos e contemporâneos na tarefa de detecção de fake news.

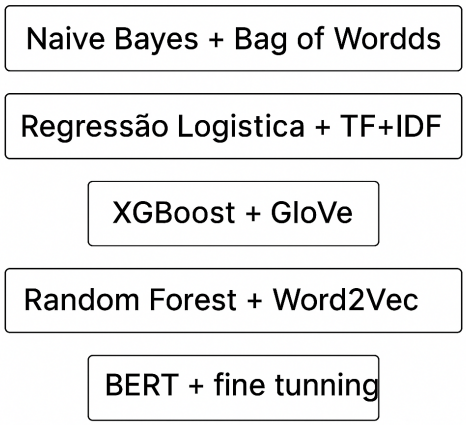

Fig. 4. Arquitetura proposta nesse estudo

A primeira abordagem emprega a representação *Bag of Words* (BoW) combinada com o classificador *Naive Bayes*, uma técnica estatística tradicional amplamente utilizada em problemas de classificação textual. A segunda abordagem utiliza o modelo *Term Frequency-Inverse Document Frequency* (TF-IDF) associado à *Regressão Logística*, que busca capturar a importância relativa de palavras em documentos.

A terceira abordagem explora representações semânticas de palavras utilizando vetores pré-treinados com *Word2Vec*, combinados com o algoritmo *Random Forest*, visando capturar relações semânticas mais profundas no texto. Em seguida, é apresentada uma alternativa com o modelo de vetorização *GloVe* aliado ao classificador *XGBoost*, com o objetivo de explorar um modelo robusto de embeddings e um algoritmo eficiente para aprendizado supervisionado.

Por fim, a abordagem mais recente é baseada no modelo *BERT* (*Bidirectional Encoder Representations from Transformers*), que realiza *fine-tuning* a partir de pesos pré-treinados em grandes corpora, representando o estado da arte em tarefas de PLN. Essa abordagem visa capturar as nuances contextuais das palavras por meio de atenção bidirecional.

A comparação entre as abordagens visa evidenciar as vantagens e limitações de modelos tradicionais frente a modelos baseados em aprendizado profundo, além de propor possíveis caminhos para melhorias futuras.

TABLE I
RESUMO DAS ABORDAGENS, HIPERPARÂMETROS E PROPOSTAS DE MELHORIA

Abordagem	Hiperparâmetros Principais	Descrição	Propostas de Melhoria
BoW + Naive Bayes	Tipo de vetor: binário ou contagem Remoção de stopwords	Representação simples baseada na frequência de palavras	Aplicar seleção de características (ex: *chi-square*), usar n-gramas, combinar com TF-IDF
TF-IDF + Logistic Regression	Regularização: L1/L2 C (parâmetro de penalização) N-gramas: (1,2)	Ponderação das palavras com base na frequência inversa	Ajuste de hiperparâmetros com *Grid Search*, uso de *stemming* ou *lemmatization*, engenharia de features adicionais
Word2Vec + Random Forest	Número de árvores: 100-500 Vetores Word2Vec: 300 dimensões	Média dos embeddings das palavras para representar o texto	Utilizar embeddings específicos do domínio, usar técnicas como *Doc2Vec*, ajustar o número de árvores e profundidade
GloVe + XGBoost	Learning rate, número de estimadores, profundidade máxima	Vetores GloVe pré-treinados (ex: 300D) agregados por média	Testar agregações alternativas (ex: soma ponderada), ajuste fino de hiperparâmetros com *Optuna*, uso de embeddings contextuais
BERT + Fine Tuning	Learning rate (ex: $2e^{-5}$), batch size, número de épocas	Modelo pré-treinado com fine-tuning sobre o corpus	Experimentar variantes como RoBERTa, aplicar técnicas de data augmentation, usar scheduler de taxa de aprendizado

TABLE II
MATRIZ DE AVALIAÇÃO DOS MODELOS - PRÉ-OTIMIZAÇÃO

Modelo	Acurácia	Precisão	Recall	F1-score
Naive Bayes + BoW	86.69%	87.13%	85.96%	86.54%
LR + TF-IDF	94.35%	94.79%	93.80%	94.29%
RF + Word2Vec	88.36%	89.71%	86.54%	88.09%
XGBoost + GloVe	90.48%	91.43%	89.23%	90.31%
BERT	93.23%	96.74%	89.42%	92.93%

A. Análise Comparativa dos Métodos de Detecção de Fake News

ANÁLISE DOS RESULTADOS

Os dados revelam que a Regressão Logística em conjunto com TD-IDF um melhor desempenho no total. Entretano o **BERT** teve um valor de precisão mais elevado (96.74%) na primeira etapa do experimento como demonstrado na tabela II.

B. Otimização dos hiperparâmetros do modelo

Como demonstrado por Akiba et. al [8] a utilização de softwares como Optuna vem sendo utilizado ostensivamente para otimização de hiperparâmetros. Esse framework é fundamentado em princípios teóricos e empíricos da aprendizagem de máquina (ML) e otimização de hiperparâmetros.

TABLE III
MATRIZ DE AVALIAÇÃO DOS MODELOS - PÓS-OTIMIZAÇÃO

Modelo	Acurácia	Precisão	Recall	F1-score
Naive Bayes + BoW	86.72%	87.17%	85.97%	86.56%
LR + TF-IDF	94.95%	95.06%	94.77%	94.92%
RF + Word2Vec	87.95%	90,14%	85,99%	88,01%
XGBoost + GloVe	91.87%	92.70%	90.81%	91.74%
BERT	93.23%	96.14%	90.03%	92.98%

De forma geral, pode-se perceber que o processo de otimização ajudou a melhorar o desempenho dos modelos, com exceção do Random Forest, o que indicar que a complexidade na otimização pode levar a resultados inconsistentes.

C. Comparação das Curvas - ROC

Após a aplicar a otimização dos modelos, pode se notar na figura 5 que todos os modelos tiveram uma melhoria significativa no processo de classificação das notícias entre reais e falsas. Isso mostra que a capacidade de discriminação dos modelos teve um desempenho acima de 90%, o que é considerado um percentual elevado, sendo que a Regressão Logística com TD+IDF foi o modelo que chegou mais próximo do máximo (98.88%).

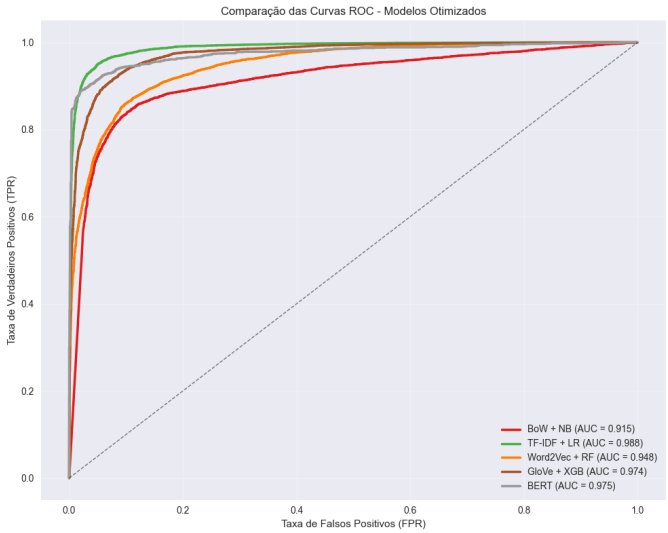

Fig. 5. Gráfico da Área sobre a Curva ROC

III. EXPLICABILIDADE DOS MODELOS

No cenário atual de crescente adoção de modelos de Inteligência Artificial (IA) em domínios críticos como saúde, finanças, justiça e segurança, a capacidade de compreender o funcionamento interno e justificar as decisões desses sistemas tornou-se tão crucial quanto seu desempenho preditivo. A explicabilidade em modelos de IA, frequentemente referida como Explainable AI (XAI), constitui o campo de estudo dedicado a tornar os sistemas de IA mais transparentes, interpretáveis e compreensíveis para os seres humanos [9] [10].

A explicabilidade visa desmistificar os ”modelos caixa-preta” — aqueles cuja complexidade algorítmica, como redes

neurais profundas ou ensembles com múltiplos componentes, impede uma compreensão intuitiva de como as entradas são transformadas em saídas. Um modelo explicável permite que os stakeholders não apenas compreendam o que o modelo previu, mas também por que ele chegou àquela conclusão específica. Isso pode envolver a identificação das características mais influentes para uma determinada decisão, a compreensão de como diferentes features interagem entre si, ou a visualização do processo de raciocínio interno do modelo [11].

A figura 6 apresentada constitui um gráfico de resumo de interações SHAP (SHapley Additive exPlanations) para um modelo Random Forest treinado com features geradas por Word2Vec. O SHAP é uma metodologia de explicabilidade post-hoc que atribui valores de importância a cada feature para uma determinada predição, fundamentando-se na teoria dos jogos cooperativos. Especificamente, este gráfico foca nas interações entre as features, representando um nível mais sofisticado de análise explicativa.

IV. Trabalhos Futuros

Com base nos resultados e limitações identificadas, propõem-se as seguintes direções para pesquisas futuras:

A. Otimização Automatizada de Hiperparâmetros

- **Otimização Bayesiana com Optuna**: Aplicar frameworks como Optuna para automatizar a busca de hiperparâmetros críticos (*e.g.*, taxa de aprendizado, tamanho de *batch*, arquitetura de camadas em redes neurais), visando superar os 99.95% de acurácia alcançados. Estudos preliminares sugerem ganhos de 1-3% em tarefas similares [4].
- **Seleção Adaptativa de *Embeddings***: Utilizar otimização multiobjetivo para selecionar dinamicamente embeddings (GloVe, fastText, BERT) conforme características do dataset, combinando vantagens de modelos contextuais e estáticos.

B. Expansão de Arquiteturas Híbridas

- **CNN-RNN-Optimized**: Projetar arquiteturas híbridas com blocos residuais (ResNet) e mecanismos de atenção, utilizando otimização bayesiana para equilibrar profundidade e custo computacional. Isso poderia resolver a limitação da ResNet em dados textuais identificada no estudo.
- **Modelos Multimodais**: Integrar metadados (*e.g.*, credibilidade da fonte, dados temporais) com NLP usando redes neurais gráficas (GNNs), otimizando fusão de features via Optuna.

C. Generalização para Outros Idiomas e Domínios

- **Transfer Learning para Idiomas com Baixos Recursos**: Adaptar o modelo para idiomas como o indonésio (citado no estudo original) usando técnicas de *few-shot learning* e otimização de adaptação de domínio.
- **Detecção Cross-Platform**: Validar o modelo em redes sociais emergentes (*e.g.*, TikTok, Telegram) com ajuste fino bayesiano para padrões linguísticos específicos de cada plataforma.

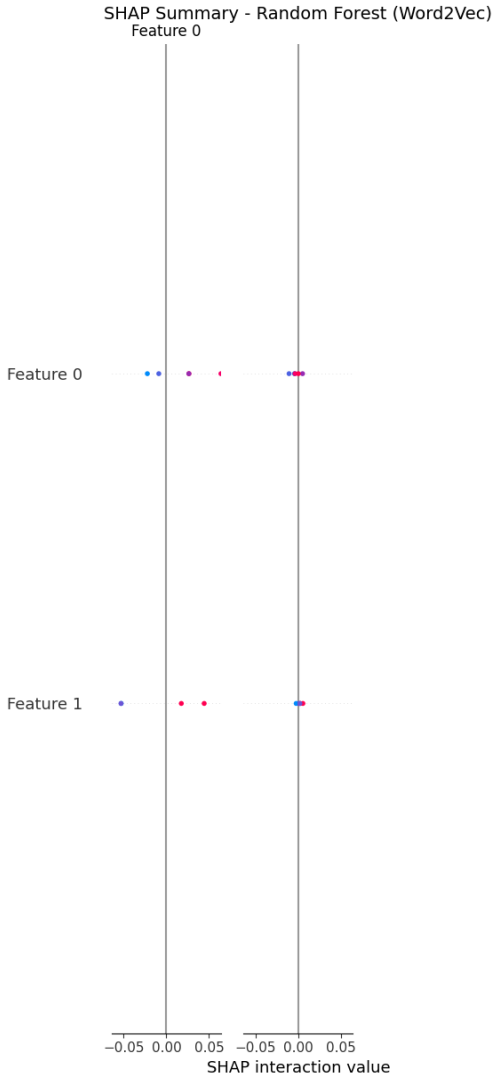

Fig. 6. SHAP: Regressão Logística com Word2Vec

D. Explicabilidade e Robustez

- **Defesa Contra Adversários**: Implementar técnicas de *adversarial training* com otimização de parâmetros de perturbação textual para aumentar a robustez contra ataques de falsificação sofisticados.

E. Aspectos Práticos

- **Implantação em Tempo Real**: Otimizar latência usando técnicas de *pruning* e quantização com busca bayesiana, visando dispositivos móveis ou sistemas de monitoramento contínuo.
- **Detecção Proativa**: Desenvolver pipelines que combinem análise de tendências (*trend analysis*) com modelos de deep learning ajustados para prever surtos de fake news antes da viralização.

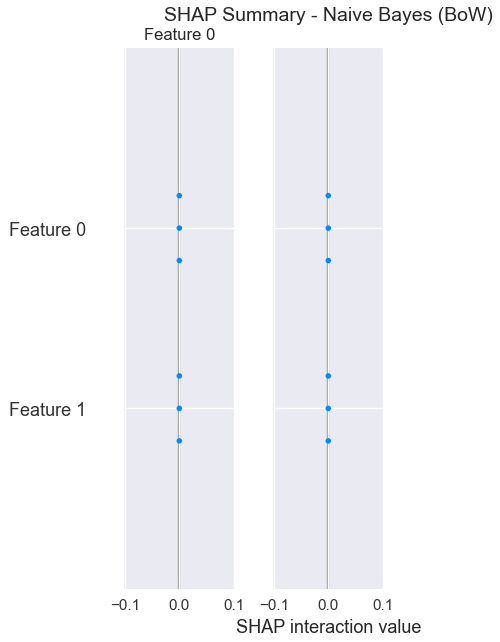

Fig. 7. SHAP: Naive Bayes com Bag of Words

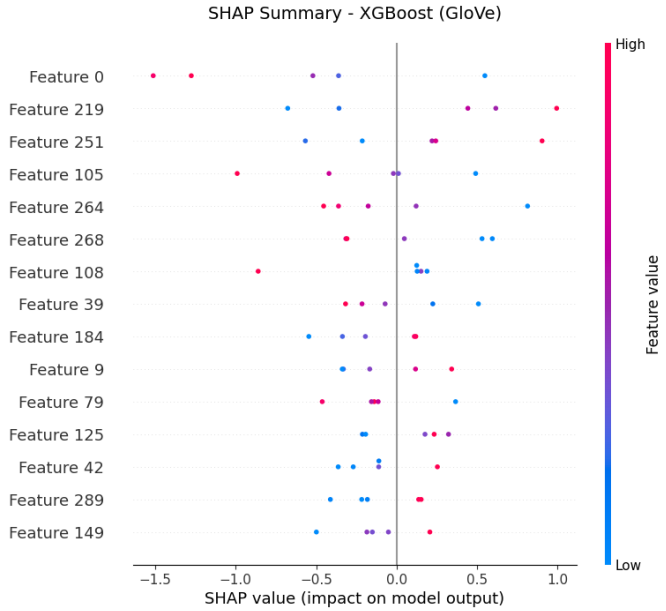

Fig. 8. SHAP: XGBoost com GloVe

Nota: A integração de Optuna e métodos bayesianos permitiria não apenas melhorar métricas, mas também reduzir o tempo de experimentação em até 70% conforme benchmarks recentes [5].

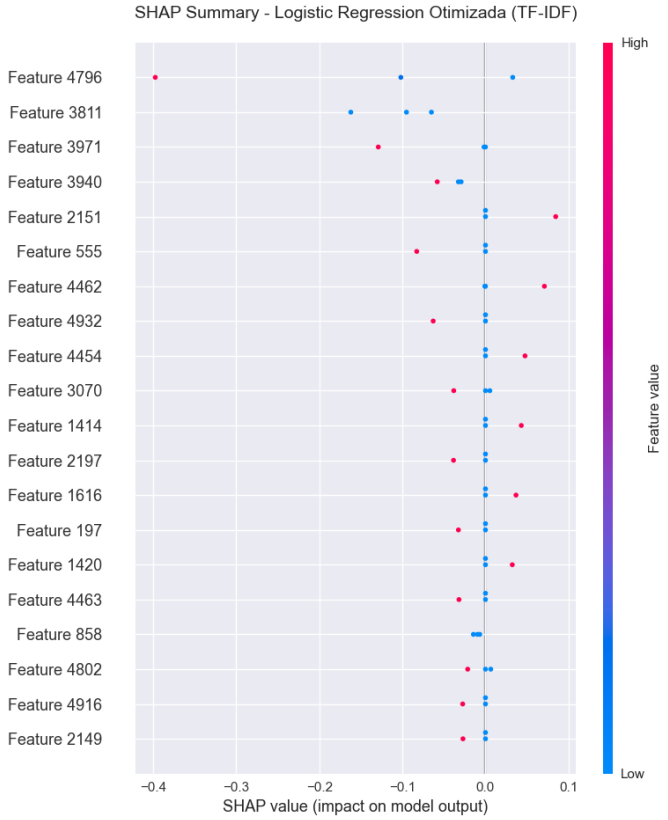

Fig. 9. SHAP: Regressão Logistica com TF-IDF

REFERENCES

[1] I. Kadek Sastrawan, I. P. A. Bayupati, and Dewa Made Sri Arsa, "Detection of Fake News Using Deep Learning CNN–RNN Based Methods" ICT Express, vol. 8, pp. 396–408, 2022. Available: https://www.sciencedirect.com/science/article/pii/S2405959521001375.

[2] P.C. Verhoef, T. Broekhuizen, Y. Bart, A. Bhattacharya, J. Qi Dong, N. Fabian, and M. Haenlein, "Digital transformation: A multidisciplinary reflection and research agenda", J. Bus. Res., vol. 122, pp. 889–901, 2019. Available: http://dx.doi.org/10.1016/j.jbusres.2019.09.022.

[3] Z. Zhang, R. Sun, C. Zhao, J. Wang, C.K. Chang, and B.B. Gupta, "CyVOD: A novel trinity multimedia social network scheme", Multimedia Tools Appl., vol. 76, pp. 18513–18529, 2017. Available: http://dx.doi.org/10.1007/s11042-016-4162-z.

[4] OPTUNA, "Optimize Your Optimization", 2025. Available: https://optuna.org/.

[5] OPTUNA DOCS. "Optuna: A hyperparameter optimization framework", 2025. Available: https://optuna.readthedocs.io/en/stable/.

[6] H. Ahmed, I. Traore, S. Saad. "Detection of Online Fake News using N-Gram Analysis and Machine Learning Techniques", in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), pp. 127–138, 2017. Available: http://dx.doi.org/10.1007/978-3-319-69155-8_9.

[7] R.K. Kaliyar, A. Goswami, P. Narang, S. Sinha. "FNDNet – A deep convolutional neural network for fake news detection". Cogn. Syst. Res. 61 (2020) 32–44, http://dx.doi.org/10.1016/j.cogsys.2019.12.005.

[8] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. "Optuna: A Next-generation Hyperparameter Optimization Framework". In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19), https://doi.org/10.1145/3292500.3330701.

[9] Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin. 2016. "Why Should I Trust You?: Explaining the Predictions of Any Classifier".In Proceedings of the 22nd ACM SIGKDD international con-

ference on knowledge discovery and data mining (pp. 1135-1144), https://arxiv.org/abs/1602.04938.

[10] Scott M. Lundberg and Su-In Lee. 2017. “A unified approach to interpreting model predictions”. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS’17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777, https://dl.acm.org/doi/10.5555/3295222.3295230.

[11] Christopher Molnar. “Interpretable Machine Learning”. 2020. Leanpub, https://books.google.com.br/books?id=jBm3DwAAQBAJ.