ESTUDO DE CASO PARA DETECÇÃO DE FAKE NEWS: INSIGHTS DE APRENDIZADO DE MÁQUINA E MODELOS DE LINGUAGEM GRANDE EM PORTUGUÊS BRASILEIRO

Laianne dos Santos Protasio[1]; Oberdan Rocha Pinheiro[2]
[1] Bolsista; Pesquisa, Desenvolvimento e Inovação – SENAI CIMATEC; protasiolaianne@gmail.com
[2] Centro Universitário SENAI CIMATEC; Salvador-BA; oberdan.pinheiro@fieb.org.br

RESUMO

A disseminação de notícias falsas representa um desafio global significativo, afetando vários setores, desde a economia até à saúde pública. Apesar da extensa pesquisa dedicada à detecção de notícias falsas, o domínio dos conjuntos de dados em língua inglesa resulta numa deficiência no desenvolvimento de estudos em contextos não ingleses, como o português brasileiro. Com isso, este estudo comparou modelos de aprendizagem de máquina, e a eficácia do modelo de linguagem BERT no discernimento de notícias falsas em português. Foi visto que, a Regressão Logística e o Classificador SDG demonstraram notável acurácia e F1-score na classificação de artigos de notícias. Além disso, o modelo BERT apresentou uma acurácia de 98%, ressaltando a sua eficácia na categorização precisa de artigos de notícias.

PALAVRAS-CHAVE: Classificação de texto; Modelos de linguagem grande (LLMs); Fake News; Aprendizado de máquina.

1. INTRODUÇÃO

A disseminação de notícias falsas é hoje um problema em todo o mundo. A facilidade com que a informação atravessa a Internet pôs em causa a confiabilidade das fontes de notícias.[1,2] Este fenômeno global está presente na vida quotidiana através dos canais das plataformas de redes sociais e publicações online. O crescimento exponencial da informação divulgada a cada dia ampliou o impacto das imprecisões, levando a repercussões em diversas esferas, como já foi relatado na influência na tomada de decisões econômicas,[3] em decisões políticas,[4,5] e em situações críticas de saúde na pandemia de COVID-19,[6,7] afetando diretamente a vida de populações.

Estes diferentes cenários destacam a importância no combate a propagação da desinformação. Visto a sua urgência, para resolver esse problema, autores realizaram diferentes abordagens para detecção de notícias falsas. Com foco em notícias de tipo jornalístico Giordani et al., realizaram também uma abordagem baseada em aprendizado de máquina utilizando técnicas de Processamento de Linguagem Natural, como TF-IDF e Word2Vec. Foi avaliado o desempenho de diferentes algoritmos de classificação, como Regressão Logística, Máquina de Vetores de Suporte, Random Forest, AdaBoost, e LightGBM. Os resultados demonstraram o alto desempenho dos modelos, principalmente do modelo LightGBM treinado em recursos TF-IDF.

O Processamento de Linguagem Natural (PLN) passou por uma fase revolucionária com o advento de modelos de transformadores pré-treinados. A arquitetura de transformador permitu o treinamento de modelos para PLN de uma forma mais eficiente. Modelos como o BERT,[9] são treinados em grandes conjuntos de dados sendo chamados como Modelos de Linguagem Grande (LLMs). Esses modelos demonstraram eficácia notável em diversas tarefas de PLN, como em classificação de texto,[10] reconhecimento de entidade nomeada (NER),[11] tradução automática,[12,13] resposta a perguntas,[14,15] geração de texto,[16,17] e sumarização.[18,19]

Apesar da extensa pesquisa dedicada à detecção de notícias falsas, a maioria dos estudos concentra-se predominantemente em notícias na língua inglesa. Isto resultou numa notável deficiência de conjuntos de dados rotulados para detecção de notícias falsas em outros idiomas. Além disso, questões cruciais neste domínio permanecem sem resposta, enfatizando a necessidade de investigações abrangentes e inclusivas para além dos limites das notícias em língua inglesa. Portanto, este estudo compara metodologias distintas empregando aprendizado de máquina e modelo de linguagem de código aberto pré-treinado (BERT), com o objetivo de avaliar seu desempenho na categorização precisa de artigos de notícias falsas em língua portuguesa brasileira.

2. METODOLOGIA

2.1. Dados

Visto a deficiência mencionada sobre a disponibilidade de conjuntos de dados rotulados, principalmente em português, Monteiro et al., apresentaram o primeiro corpus de referência na área para a língua portuguesa. O corpus chamado Fake.Br, é composto por notícias verdadeiras e falsas alinhadas,

atendendo a estudos linguísticos e propósitos de aprendizado de máquina. O processo envolveu uma recolha manual e verificação de 7.200 artigos de notícias, abrangendo um período de dois anos, de Janeiro de 2016 a Janeiro de 2018.

A coleção se concentrou em quatro sites (Diário do Brasil, A Folha do Brasil, The Jornal Brasil e Top Five TV), com atenção à filtragem de meias verdades de notícias falsas. Notícias verdadeiras foram coletadas de forma semiautomática usando um rastreador da web, utilizando palavras-chave e medidas de similaridade lexical. O corpus abrange diversos assuntos, categorizados em política, TV e celebridades, sociedade e notícias diárias, ciência e tecnologia, economia e religião.

Para realizar efetivamente a detecção de Fake News, é necessário a aplicação de técnicas de pré-processamento nos dados, quando é feita a utilização dos modelos clássicos utilizados neste trabalho descritos nos tópicos seguintes. Assim, após o download dos textos, foi criada e rotulada a base de dados em arquivo csv. As notícias verdadeiras foram classificadas como Falsas (zero) e as notícias fake como True (um). O primeiro momento foi de pré-processamento e vetorização dos dados textuais usando embeddings Word2Vec. A implementação começa importando as principais bibliotecas como Gensim para embeddings Word2Vec e outras bibliotecas padrão como pandas, NumPy e NLTK para manipulação de dados e funcionalidades de processamento de linguagem natural.

O modelo Word2Vec pré-treinado é carregado no script. Este modelo foi previamente treinado em um grande corpus e encapsula relações semânticas entre palavras em forma vetorizada. Em seguida, os dados do texto passam por pré-processamento, envolvendo letras minúsculas, remoção de caracteres não alfabéticos e eliminação de stopwords (palavras comuns com pouco valor semântico). O modelo Word2Vec é então empregado para obter representações vetoriais para cada documento, calculando o vetor médio das palavras constituintes.

Os vetores resultantes são transpostos e organizados em um DataFrame. Este DataFrame é então concatenado com o conjunto de dados original, que inclui o texto bruto, o texto pré-processado sem palavras irrelevantes e os rótulos de destino.

O DataFrame final pré-processado e vetorizado é salvo em um arquivo binário para uso futuro no treinamento dos modelos clássicos.

2.2. Modelos

Após o pré-processamento, neste estudo foi usado algumas técnicas de PLN. Como Regressão Logistica, Classificador SDG, Classificador de árvore de decisão e Classificador Kneighbors.

As etapas seguintes constituíram-se na avaliação, comparação do desempenho dos modelos e ajuste de hiperparâmetros. Para isso foi feita a validação cruzada com os modelos sklearn tradicionais, sendo gerada métricas de avaliação para a comparação dos modelos.

O modelo de linguagem pré-treinado utilizado neste estudo foi o BERT (*Bidirectional Enconder Representations from Transformers*). [9] É um modelo de processamento de linguagem natural que avançou significativamente no campo do aprendizado de máquina e da compreensão da linguagem. Desenvolvido pelo Google em 2018, o BERT se tornou um dos modelos mais influentes e amplamente utilizados em tarefas de PLN.

A etapa inicial para o desenvolvimento do estudo, envolveu a configuração do ambiente Python necessário e a instalação importação de bibliotecas principais. Em seguida, é feito o carregamento dos dados, a atribuição dos rótulos e construção do DataFrame. O conjunto de dados é então dividido em treinamento, teste e validação, garantindo uma distribuição equilibrada dos dados para treinamento e avaliação do modelo. É criada uma estrutura ‘DatasetDict’, necessária pela compatibilidade com a biblioteca Transformers, o que facilita a integração do código com as funcionalidades dos transformadores. A proxima etapa concentrou-se na tokenização, sendo empregado a classe ‘AutoTokenizer’ para carregar o tokenizer para o modelo BERT pré-treinado ("bert-base-uncased").

Com os dados preparados e tokenizados, a fase final concentrou-se no treinamento e avaliação do modelo. Isso envolve configurar o modelo, definir parâmetros de treinamento e utilizar os recursos da GPU.

3. RESULTADOS E DISCUSSÃO

Na avaliação dos quatro modelos de aprendizado de máquina – Regressão Logística, Classificador SDG, Classificador de Árvore de Decisão e Classificador Kneighbours – conduzida neste estudo, foram observadas métricas de desempenho notáveis para Regressão Logística e Classificador SDG, particularmente em termos de acurácia e F1-score. Os resultados detalhados para métricas de acurácia, precisão, recall e F1-score para todos os modelos são apresentados na Tabela 1.

Em contraste com descobertas já vistas,[21] nosso estudo ressalta a eficácia da Regressão Logística e do Classificador SDG no discernimento entre artigos de notícias falsos e verdadeiros. Nosso estudo

identificou a Regressão Logística como um modelo que demonstra um bom desempenho, desafiando a noção de que pode não ser eficaz para tal classificação. Esta diferença enfatiza a natureza dinâmica do desempenho do modelo em diferentes conjuntos de dados e contextos, acentuando a importância de uma melhor avaliação em cenários específicos. Nossos resultados apresentam uma perspectiva alternativa, contribuindo para o discurso contínuo sobre a adequação de vários algoritmos de aprendizado de máquina para tarefas de classificação de notícias.

Tabela 1. Valores encontrados para as métricas avaliadas em cada algoritmo.

Algoritmos	*Acurácia*	*Precisão*	*Recall*	*F1-Score*
Regressão Logistica	86%	88%	83%	85%
SDG	85%	93%	90%	85%
Árvore de decisão	76%	80%	80%	75%
Kneighbors	64%	97%	33%	48%

Sobre a eficácia do modelo BERT para classificar artigos de notícias verdadeiros e falsos, surgiram resultados convincentes durante a avaliação do modelo. A acurácia atingiu 98%, ressaltando a excepcional capacidade do modelo de classificar com precisão artigos de notícias em suas respectivas categorias. Além disso, a perda observada, de 17%, enfatiza ainda mais a proficiência do modelo em minimizar a diferença entre os rótulos previstos e reais. Os resultados da avaliação, resumidos na Tabela 2, destacam o desempenho robusto do modelo BERT no conjunto de validação.

Tabela 2. Avaliação do modelo BERT.

Parâmetros	*Valores*
Perda (eval_loss)	0.17
Acurácia (eval_accuracy)	0.98
Tempo de execução (eval_runtime)	11.6
Amostras por segundo (eval_samples_per_second)	92.55
Passos por segundo (eval_steps_per_second)	11.5
Épocas (epoch)	10

Além disso, a eficiência demonstrada pelo modelo durante a fase de avaliação, com notável velocidade de previsão de novos dados, ressalta sua aplicabilidade prática em cenários em tempo real. Este aspecto é particularmente significativo para aplicações que exigem tomadas de decisão rápidas com base na classificação de artigos noticiosos.

4. CONSIDERAÇÕES FINAIS

No nosso estudo, a Regressão Logística e o Classificador SDG demonstraram notável acurácia e F1-score na classificação de artigos de notícias. Além disso, o modelo BERT apresentou uma acurácia de 98%, ressaltando a sua eficácia na categorização precisa de artigos de notícias. Nossas descobertas ressaltam a habilidade do modelo BERT na classificação de artigos de notícias, revelando sua alta acurácia, baixa perda e velocidade preditiva eficiente. Recomenda-se a investigação contínua ao longo de várias épocas para obter insights mais profundos sobre a trajetória de aprendizagem do modelo e o desempenho geral em diversos conjuntos de dados.

5. REFERÊNCIAS

[1] PÉREZ-ESCODA, A., Pedrero-Esteban, L. M., Rubio-Romero, J., and Jiménez-Narros, C. **Fake news reaching young people on social networks:** Distrust challenging media literacy. Publications, v. 9, n. 2, p. 24, 2021.
[2] SHU, K., Sliva, A., Wang, S., Tang, J., and Liu, H.. **Fake news detection on social media:** A data mining perspective. ACM SIGKDD explorations newsletter, v. 19, n. 1, p. 22-36, 2017.
[3] AUSAT, Abu Muna Almaududi. **The Role of Social Media in Shaping Public Opinion and Its Influence on Economic Decisions**. Technology and Society Perspectives (TACIT), v. 1, n. 1, p. 35-44, 2023.
[4] TURCILO, Lejla; OBRENOVIC, Mladen. **Misinformation, disinformation, malinformation:** Causes, trends, and their influence on democracy. Heinrich Böll Foundation, 2020.
[5] GALEOTTI, A. et al. **Political disinformation and voting behavior:** Fake news and motivated reasoning. Notizie di Politeia, v. 37, n. 142, p. 64-85, 2021.
[6] ZHENG, Han; WANG, Xiaohui; HUANG, Yi-Hui. **Fake news in a time of plague:** Exploring individuals' online information management in the COVID-19 era. Computers in Human Behavior, v. 146, p. 107790, 2023.
[7] HADLINGTON, Lee et al. **Perceptions of fake news, misinformation, and disinformation amid the COVID-19 pandemic:** A qualitative exploration. Psychology of Popular Media, v. 12, n. 1, p. 40, 2023.
[8] GIORDANI, Luiz et al. **Fake News BR:** A Fake News Detection Platform for Brazilian Portuguese. arXiv preprint arXiv:2309.11052, 2023.
[9] DEVLIN, Jacob et al. **Bert:** Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
[10] WANI, Apurva et al. **Evaluating deep learning approaches for covid19 fake news detection**. Springer International Publishing, 2021.
[11] SOUZA, Fábio; NOGUEIRA, Rodrigo; LOTUFO, Roberto. **Portuguese named entity recognition using BERT-CRF**. arXiv preprint arXiv:1909.10649, 2019.
[12] BARRIERE, Valentin; BALAHUR, Alexandra. **Improving sentiment analysis over non-english tweets using multilingual transformers and automatic translation for data-augmentation**. arXiv preprint arXiv:2010.03486, 2020.
[13] CAMGOZ, Necati Cihan et al. **Sign language transformers:** Joint end-to-end sign language recognition and translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020.
[14] RADFORD, Alec et al. **Language models are unsupervised multitask learners**. OpenAI blog, v. 1, n. 8, p. 9, 2019.
[15] QU, Chen et al. **BERT with history answer embedding for conversational question answering**. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, 2019.
[16] MAGER, Manuel et al. **GPT-too:** A language-model-first approach for AMR-to-text generation. arXiv preprint arXiv:2005.09123, 2020.
[17] TOPAL, M. Onat; BAS, Anil; VAN HEERDEN, Imke. **Exploring transformers in natural language generation:** Gpt, bert, and xlnet. arXiv preprint arXiv:2102.08036, 2021.
[18] GARG, Apar et al. **NEWS article summarization with pretrained transformer**. In: Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I 10. Springer Singapore, 2021.
[19] KHANDELWAL, Urvashi et al. **Sample efficient text summarization using a single pre-trained transformer**. arXiv preprint arXiv:1905.08836, 2019.
[20] MONTEIRO, Rafael A. et al**. Contributions to the study of fake news in portuguese:** New corpus and automatic detection results. Springer International Publishing, 2018.
[21] SUTRADHAR, Biplob Kumar et al. **Machine Learning Technique Based Fake News Detection**. arXiv preprint arXiv:2309.13069, 2023.