Classificação de Notícias em Português Utilizando Modelos Baseados em Transferência de Aprendizagem e Transformers

Wagner Narde[1], João Mendanha[2], Henrique Barbosa[4], Frederico Coelho[4], Bruno Santos[3], Luiz Torres[2]

[1]Grupo Energisa – Brasil

[2]Dept. de Computação e Sistemas – Universidade Federal de Ouro Preto (UFOP)

[3]Dept. de Ciência da Computação – Universidade Federal da Bahia (UFBA)

[4]Dept. de Engenharia Eletrônica – Universidade Federal de Minas Gerais (UFMG)

wagner.b.n@hotmail.com, bruno.ps@ufba.br, luiz.torres@ufop.edu.br

Abstract. *Fake news spreads faster on some social networks than regular news, which can have different ramifications, from influences on election outcomes to deaths due to incorrect treatments of diseases. This work aims to employ methods based on transfer learning and Transformer-based machine learning models to classify the veracity of tweets in the Portuguese language (Brazil pt-BR). To this aim, a reliably labeled database was created and opened to free access. The dataset relates posts extracted from X (formerly Twitter) and their proximity with facts or fake information. Five Transformer models were subsequently trained in Portuguese. The fine-tuned BERT model, initialized with pre-training on Portuguese text, achieved superior performance, yielding an accuracy of 95.1%.*

Resumo. *Fake news se espalha mais rápido em algumas redes sociais do que notícias regulares, o que pode ter diferentes consequências, desde influências nos resultados eleitorais até mortes devido a tratamentos incorretos de doenças. Este trabalho tem como objetivo empregar métodos baseados em aprendizado por transferência e modelos de aprendizado de máquina baseados em Transformers para classificar a veracidade de tweets na língua portuguesa (Brasil pt-BR). Para isso, foi criada uma base de dados confiável e rotulada, aberta para acesso gratuito. O conjunto de dados relaciona postagens extraídas do X (anteriormente conhecido como Twitter) e sua proximidade com fatos ou informações falsas. Subsequentemente, cinco modelos Transformer foram treinados em português. O modelo BERT ajustado, inicializado com pré-treinamento em textos em português, alcançou um desempenho superior, obtendo uma acurácia de 95.1%.*

1. Introdução

Os veículos de comunicação de grande circulação por muito tempo foram jornais, revistas, rádio e televisão. Hoje, notícias circulam através de vídeos no *YouTube*, portais de notícia e em redes sociais como *Facebook*, X (Antigo *Twitter*) e *WhatsApp*. Tornando a Internet, um dos principais meios de comunicação e consumo de notícias. No Brasil, 65% usam a Internet e suas aplicações como principais fontes de informação, nos Estados

Unidos 53% e, no mundo, o número estimado é de 62% [NegociosSC 2024, Gente 2024, DataReportal 2024, Data 2024]. O que, por um lado, demonstra que a Internet ampliou o acesso à informação, mas, por outro lado, também transformou a forma como as notícias são consumidas e compartilhadas.

A ascensão da Internet e, consequentemente, das redes sociais democratizaram a produção de notícias, permitindo que qualquer pessoa assuma o papel de produtor de conteúdo sem a supervisão tradicional de jornalistas. Este fenômeno pode ter impactado negativamente na qualidade das informações disseminadas, resultando em um aumento de notícias que propagam desinformação ou divulgam informações falsas [Reis et al. 2019, Vargas et al. 2021].

2. Proposta de Modelo

Neste trabalho, propomos um modelo baseado em transferência de aprendizagem, *transformers* e aprendizagem supervisionada para classificar textos em português nas redes sociais, com foco na plataforma X. Também criamos uma base de dados em português[1] (162 amostras e balanceada), que relaciona textos da plataforma X com sua veracidade, visando melhorar a detecção de *fake news* e promover a qualidade da informação nas redes sociais.

3. Metodologia

O primeiro passo foi a construção de uma base de dados contendo postagens de usuários da rede social X (Antigo *Twitter*). Foi utilizada a ferramenta *Get Old Tweets (GOT)* [Henrique 2018] para coletar *tweets* históricos, incluindo notícias falsas. Dessa forma, cada *tweet* foi analisado e classificado manualmente para determinar sua proximidade com o fato, assegurando que o conjunto de dados fosse rigoroso e preciso. Esse processo permitiu a criação de um conjunto de dados robusto e rotulado, essencial para o treinamento e validação eficazes do modelo de classificação de textos em português proposto. Em seguida, os textos coletados passam por uma fase de preparação, onde são inicialmente pré-processados e, posteriormente, rotulados. Após esse processo, os dados são ajustados para servirem como entradas adequadas para os modelos de aprendizagem de máquina.

3.1. Conjunto de dados: Coleta e Processamento

Este trabalho utiliza dados textuais em português extraídos da plataforma de rede social X. A plataforma permite a extração de informações através de sua *Application Programming Interface* (API). Com a rede social selecionada, iniciou-se a coleta de dados para compor a base de dados. O processo de obtenção dos *tweets* consistiu em buscar no site de checagem de notícias verdadeiras ou falsas (LUPA[2]) e pesquisar por elas utilizando a ferramenta GOT [Henrique 2018]. Para isso, foram realizadas filtragens de notícias e alinhamento temporal aproximado para obtenção das postagens realizadas sobre a notícia verificada.

[1]https://github.com/WagnerNarde/ML-Transformers-Tweets-falsos
[2]https://lupa.uol.com.br/

3.2. Seleção dos Modelos de Aprendizagem

Neste trabalho, foram adotados modelos de aprendizagem baseados em *Transformers*. Originalmente desenvolvido para tradução automática, o *Transformer* se destacou por sua capacidade de capturar relações de dependência de longo alcance de forma eficaz. Buscou-se por modelos *Transformers* que receberam pré-treinamento em português, visando aproveitar a transferência de aprendizagem. Como resultado, optou-se por ajustar os seguintes modelos: BERT base pré-treinado em português brasileiro por [Souza et al. 2020]; BERT base pré-treinado em 104 idiomas, incluindo português, por [Devlin et al. 2019]; RoBERTa pré-treinado por [Liu et al. 2019] com um *corpus* de 6,9 milhões de frases em português; XLM-R base, pré-treinado por [Conneau et al. 2020] em 100 idiomas, incluindo português; e, por fim, o modelo ELECTRA uncased [Clark et al. 2020], pré-treinado especificamente em português.

3.3. Treinamento

Para avaliar a capacidade de generalização do modelo, foi utilizado o método de validação cruzada com 10 partições (10-fold cross-validation). Este método divide o conjunto de dados em 10 sub-conjuntos. Cada subconjunto é usado uma vez como conjunto de teste, enquanto os restantes são usados como conjunto de treinamento. Esse processo é repetido 10 vezes, garantindo que cada amostra do conjunto de dados seja utilizada para testes ao menos uma vez. Esse procedimento não apenas melhora a capacidade de generalização do modelo, mas também fornece uma estimativa mais robusta do desempenho do modelo em dados não vistos.

4. Resultados

As métricas de avaliação incluem Acurácia, F1-Score, Precisão, Sensibilidade (Recall) e MCC.

Tabela 1. Resultados obtidos em cada modelo

Modelo	Épocas	Acurácia	F1	Precisão	Sensibilidade	MCC
ELECTRA (uncased)	10	0.864	0.848	0.883	0.824	0.720
RoBERTa pré-treinado em Português	7	0.901	0.897	0.852	0.962	0.812
XLM-R pré-treinado em multi-idiomas	9	0.903	0.898	0.883	0.922	0.804
BERT com Pré-treinamento em Português	6	0.944	0.955	0.944	0.971	0.887
BERT com Pré-treinamento Multi-idioma	10	0.914	0.918	0.900	0.944	0.825

Os resultados deste trabalho, apresentados na Tabela 1, mostram que cada modelo de aprendizado de máquina treinado para a classificação de notícias em português teve um desempenho variado, dependendo do número de épocas e das características do próprio modelo. O ELECTRA (*uncased*), treinado com 10 épocas, apresentou o desempenho mais baixo, o que pode ser atribuído à falta de diferenciação entre letras maiúsculas e minúsculas, bem como à qualidade dos pesos de pré-treinamento disponíveis. O modelo RoBERTa pré-treinado em Português, configurado com 7 épocas, superou o ELECTRA, beneficiando-se de uma arquitetura que captura melhor as nuances linguísticas do português e diferencia entre maiúsculas e minúsculas. O modelo XLM-R pré-treinado em multi-idiomas, com 9 épocas, demonstrou uma leve superioridade em relação ao RoBERTa em termos de acurácia e F1, aproveitando o conhecimento adquirido em múltiplos idiomas para melhorar a compreensão semântica. Já o BERT com Pré-treinamento

Multi-idioma, utilizando 10 épocas, mostrou robustez com acurácia e F1 acima de 0.9, destacando-se pela capacidade de transferir conhecimento linguístico de um corpus multilingue para o português. Por fim, o BERT pré-treinado em Português foi o modelo com melhor desempenho geral, utilizando apenas 6 épocas de treinamento. Este modelo se destacou na classificação correta das notícias, com acurácia, F1 e MCC superiores, evidenciando a eficácia do pré-treinamento específico em português e a importância do ajuste fino dos hiperparâmetros para maximizar a eficácia do modelo em tarefas específicas de classificação de texto.

5. Discussão

Os modelos ELECTRA *uncased* pré-treinado em Português e RoBERTa pré-treinado em Português apresentaram resultados abaixo do esperado, pode-se levantar a questão de que se tais modelos passaram pelo mesmo processo de pré-treinamento dos outros métodos. O modelo RoBERTa exige mais recursos computacionais comparado com o Bert, além de ser um aprimoramento do mesmo, portanto, melhores resultados eram esperados desse modelo. O modelo ELECTRA sendo um modelo *uncase*, esperava-se um desempenho abaixo dos outros classificadores pré-treinados exclusivamente em português. Ainda assim, acredita-se que o modelo não conseguiu generalizar bem o problema.

O XLM-R foi um modelo originalmente proposto para a tradução de idiomas, por isso ele está disponível em versão multi-idiomas pré-treinado em vários idiomas, inclusive português. Apesar do XLM-R não ter sido originalmente proposto para classificação de texto, ele obteve resultados melhores que o ELECTRA.

O Modelo BERT com Pré-treinamento em Português obteve acurácia e F1 superiores a todos os outros modelos, mostrando que o pré-treinamento em português feito por [Souza et al. 2020] foi muito eficiente e contribuiu positivamente para o bom desempenho do modelo. Os resultados preliminares mostraram que o modelo foi capaz de classificar noticias de uma base de dados relativamente pequena, bases de dados com poucas amostras é um desafio em algumas áreas, como na saúde.

6. Conclusões

Este trabalho apresentou uma abordagem para detecção de *tweets* falsos em português através de NLP. Além disso, foi criada e disponibilizada uma base de dados balanceada com *tweets* classificados de forma confiável. A base possibilitou o treinamento de modelos para detecção de notícias falsas. Sendo que o modelo BERT com 6 épocas foi o melhor comparado aos outros modelos testados.

7. Trabalhos Futuros

Na continuação do trabalho, pretendemos estender a avaliação comparativa com outros modelos estado da arte da literatura de classificação de texto baseados em aprendizado profundo. Pretende-se aumentar a base de dados com mais dados rotulados, mantendo a confiabilidade, e também buscar dados de outras fontes. Além de mostrar os resultados da classificação de notícias verdadeiras, planeja-se apresentar também os resultados de classificação das notícias falsas, assim como utilizar outras estratégias para o treinamento, como a validação cruzada com 5 partições.

Referências

[Clark et al. 2020] Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

[Conneau et al. 2020] 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 representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

[Data 2024] Data, P. (2024). Global social media users in 2024. Accessed: 2024-06-28.

[DataReportal 2024] DataReportal (2024). Social media users 2024 (global data & statistics). Accessed: 2024-06-28.

[Devlin et al. 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[Gente 2024] Gente, G. (2024). Pandemia e o consumo de notícias nas redes sociais. `https://gente.globo.com/pandemia-e-o-consumo-de-noticias-nas-redes-sociais/`. Acessado em 28 de junho de 2024.

[Henrique 2018] Henrique, J. (2018). Get old tweets programatically. Repository on GitHub.

[Liu et al. 2019] 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 preprint arXiv:1907.11692*.

[NegociosSC 2024] NegociosSC (2024). O uso da internet, redes sociais e mídia no brasil em 2024. `https://www.negociossc.com.br/blog/o-uso-da-internet-redes-sociais-e-midia-no-brasil-em-2024/`. Acessado em 28 de junho de 2024.

[Reis et al. 2019] Reis, J. C. S., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81.

[Souza et al. 2020] Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

[Vargas et al. 2021] Vargas, F., Benevenuto, F., and Pardo, T. (2021). Toward discourse-aware models for multilingual fake news detection. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 210–218.