Estudo Comparativo entre um Algoritmo de Evolução Diferencial e um algoritmo Genético para classificação de *Fake News* na Web

Vinícius Maciel Chaves[1], **Maria Laura Pezzin Silva**[1], **Felipe Orlando Brum**[1], **João Gabriel Rocha Silva** [1]

[1]Instituto Federal de Mato Grosso – Campus Pontes e Lacerda (IFMT)
Caixa Postal 99 – 78.250-000 – Pontes e Lacerda – MT – Brasil

{v.chaves,m.pezzin,f.orlando,}@gmail.com,joao.gabriel@ifmt.edu.br

Abstract. *The popular fake news are false or fallacious news whose objective is to disseminate untrue information about a certain subject. These news have influenced scenarios of important meanings. In this context, it is of great interest to the scientific community to use and combat them. Different works in the field of computing were applied to this problem. Thus, the objective of this work is to develop a differential evolution algorithm that adjusts parameter weights of components included in a news item in order to classify it correctly. The intuitiveness of the algorithm is to classify in a more assertive way when compared to other work in the literature that uses the Genetic Algorithm for weighting. Results indicate that the differential evolution algorithm, for evaluation databases, is slightly more advanced in identifying and classifying these news items.*

Resumo. *As populares fake news são notícias falsas ou falaciosas cujo objetivo consiste em disseminar informações inverídicas sobre um determinado assunto. Essas notícias tem influenciado cenários de importante relevância. Nesse contexto, trabalhos no âmbito da computação foram aplicados a esse problema. Assim, o objetivo deste trabalho é desenvolver um algoritmo de evolução diferencial que ajuste pesos de parâmetros de componentes estruturais de uma notícia a fim de classificá-la corretamente. O intuito do algoritmo é classificar de uma maneira mais assertiva quando comparado a outro trabalho da literatura que utiliza o Algoritmo Genético para a ponderação. Resultados indicam que o algoritmo de evolução diferencial, para base de dados avaliada, apresenta, de forma tênue, mais êxito na identificação e classificação dessas notícias.*

1. Introdução

As *Fake News* são notícias intencionalmente falsas ou enganosas, muitas vezes criadas para espalhar desinformação ou mudar a opinião pública. Nos últimos anos, as mesmas ganharam muita notoriedade devido aos problemas causados por elas, principalmente durante as eleições, nas quais os candidatos utilizam dessas práticas para conseguir algum benefício sobre os demais. Um dos casos mais conhecidos é o do Donald Trump, que durante sua candidatura espalhou falsas notícias sobre a adversária Hillary Clinton [Nathaniel 2017].

Visto que esse problema atinge esfera mundial, existe relevância, interesse e necessidade da contribuição da comunidade científica no sentido de implementação de medidas que o solucionem [Biasi et al. 2022]. Este trabalho apresenta uma pesquisa no âmbito da identificação e classificação dessas notícias baseados em dois trabalhos: [Ferreira et al. 2020] e [Almeida et al. 2021].

Em [Ferreira et al. 2020] foi apresentado um modelo matemático capaz de classificar uma notícia como verdadeira ou falsa baseado em critérios estruturais de uma notícia, sendo eles: (1) existência de autor, (2) título de notícia em caixa alta, (3) nota do PageRank do site, (4) posição do site que veicula a notícia analisada no ranking do Google, (5) quantidade de notícias similares, (6) média do PageRank das notícias similares, (7) média das posições dos sites que vinculam as notícias similares. O modelo obteve êxito em sua proposição inicial e conseguiu classificar de modo correto grande parte das notícias.

[Almeida et al. 2021] continuaram o trabalho de [Ferreira et al. 2020] de apresentar, por meio de variações do peso dos parâmetros estruturais e pequenas alterações no modelo, uma melhor classificação de notícias como falsas ou não. O trabalho atingiu uma acurácia cerca de 5%. No projeto, a variação dos pesos foi realizada com auxílio de um Algoritmo Genético (AG), uma técnica de otimização baseada na evolução biológica. Usa uma população de soluções candidatas representadas por cromossomos e usa seleção, crossover e mutação para criar novas soluções ao longo de várias gerações até encontrar uma boa solução.

Em vista disso, neste trabalho desenvolvemos um algoritmo de Evolução Diferencial (ED) para ponderar os pesos dos critérios e, baseado na inferência deste algoritmo trabalhar melhor em um universo de números reais quando comparado ao AG, obter melhores classificações quando comparado ao algoritmo anterior.

2. Materiais e Métodos

Iniciando o percurso metodológico, criamos uma nova base de dados manualmente, para ajuste dos parâmetros e outra para teste dos valores encontrados. A primeira com 50 notícias, sendo 25 falsas e 25 verdadeiras e a segunda com 100 notícias, sendo 50 falsas e 50 verdadeiras.A criação de uma nova base baseou-se no fato do trabalho de [Almeida et al. 2021] utilizar de apenas 50 notícias no processo de validação.

O processo de criação da base foi realizado por meio de *scripts* na linguagem de programação *Python* que, dado a URL de uma notícia, valores para os conteúdos estruturais elencados sejam extraídos de modo automático.

de modo seguinte, desenvolvemos um algoritmo de Evolução Diferencial Clássico, apresentado em [Storn and Price 1997], trata-se de um algoritmo de otimização que, assim como o AG, varia soluções por meio de operadores genéticos e, a cada geração busca apresentar uma melhor solução para o problema. A escolha por este algoritmo foi baseada no fato do algoritmo trabalhar com soluções no domínio dos números reais [Oliveira et al. 2006].

O método de classificação de uma notícia como verdadeira ou falsa segue a equação proposta em [Ferreira et al. 2020] descrita pela Equação 1. Nela, a avaliação de uma notícia é realizado por meio da multiplicação do peso de cada critério (P_i) pelo valor deste critério estrutural na notícia (C_i), um detalhamento sobre a obtenção dos valores dos

critérios é apresentado no trabalho inicial. Quando o valor da avaliação (A_i) é maior que 0.6 (60%) a notícia é classificada pelo modelo como verdadeira e caso contrário, falsa.

$$A_i = \sum_{i=1}^{N} P_i C_i \tag{1}$$

Para a execução do ED utilizamos a mesma função objetivo no processo de calibração dos parâmetros: Equação 2. A equação baseia-se na maximização da diferença entre a soma dos valores da classificação das notícias verdadeiras (M) subtraído a soma dos valores de classificação das notícias falsas (Q). Assim, quanto mais alta a classificação das notícias verdadeiras e mais baixa a classificação das falsas, maior a qualidade da solução encontrada.

$$max \left(\sum_{i=1}^{M} P_i C_i - \sum_{j=1}^{Q} P_j C_j \right) \tag{2}$$

O algoritmo ED desenvolvido aplicado às 50 notícias de ajuste foram executados 30 vezes, pela lei estocástica do algoritmo, e por 100 gerações e com 100 indivíduos, valores de gerações similares aos valores utilizados em [Almeida et al. 2021] visando uma comparação honesta entre os algoritmos. Os parâmetros de operadores genéticos do ED utilizados foram 1,2 para a mutação e 0,5 para o cruzamento.

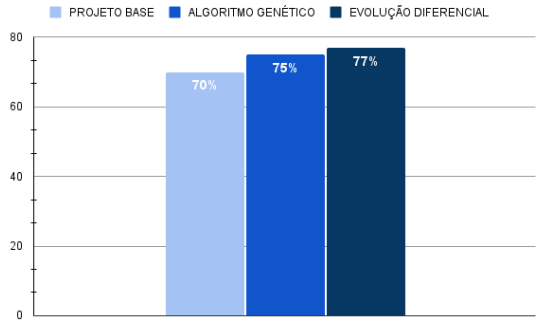
Por fim, selecionamos os melhores conjuntos de parâmetros encontrados pelo ED e aplicamos à base montada com 100 notícias visando uma classificação mais acertiva quando comparado aos demais trabalhos. O extrator de valores dos critérios a partir da URL da notícia, a busca por notícias similares e o algoritmo de Evolução Diferencial foram implementados na linguagem de programação *Python*, com auxílio das bibliotecas: urllib.request, requests, search, difflib, urlparse, sys, json.

3. Resultados e Discussões

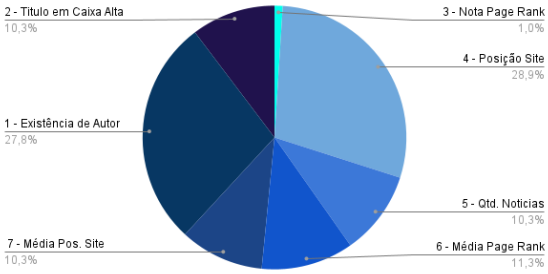
A Figura 1 apresenta os resultados obtidos pelo projeto, em 1(a) as taxas de acerto entre as classificações do modelo inicial [Ferreira et al. 2020] e da utilização do AG [Almeida et al. 2021] e o ED (nosso trabalho) e em 1(b) a ponderação entre os critérios encontrados pela melhor execução do algoritmo ED desenvolvido.

Analisando a Figura 1(a) percebe-se que o algoritmo desenvolvido se portou de forma mais eficaz em comparação ao modelo de trabalho original e que o AG. O aumento na eficácia de acertos foi de 77%, comparado aos 75% atingidos pelo AG e 70% comparado ao modelo base.

Na Figura 1 (b) é possível observar que o critério 4, nota da *Posição do site* que veicula a notícia e o critério 1, que é a existência de um autor, ou seja, notícias assinadas por um portal ou um jornalista, são os critérios que mais contribuíram e influenciaram na classificação das notícias, o terceiro critério que mais contribuiu para a classificação é a métrica baseada na quantidade e qualidade de links que um site recebe, *Média do Page Rank*, de modo análgo ao trabalho utilizado como referência [Almeida et al. 2021] que, para a base de dados testada, apresenta esses três critérios também com alta influência na classificação.



(a) Taxa de Acerto nas classificações pelos modelos



(b) Melhor indivíduo encontrado pelo algoritmo ED proposto

Figura 1. Principais resultados obtidos no findar do percurso metodológico

4. Conclusão

Este trabalho consistiu no desenvolvimento e na proposição de um estudo comparativo entre um algoritmo genético e um algoritmo de evolução diferencial para o problema de classificação das *fake news* na Web. Nosso trabalho atingiu índices positivos para proposta e alcançou uma melhor classificação quando comparada a trabalhos anteriores.

Trabalhos futuros consistem em testarmos outros algoritmos e na criação de uma plataforma Web aberta ao público para alimentação da base e visualização da classificação das notícias em tempo real. Assim, aumentando o conjunto de dados e disponibilizando a ferramenta à sociedade.

Referências

Almeida, A. L., Carrara, G., Prates, I., Nascimento, L. C., Souza, P. H., Almeida, T., Cani, R., and Silva, J. G. (2021). Modelo matemático apoiado por um algoritmo genético para classificação de fake news na web. In *Anais do VIII Encontro Nacional de Computação dos Institutos Federais*, pages 17–20, Porto Alegre, RS, Brasil. SBC.

Biasi, M. F., Amorim, M. M. R., and Katz, L. (2022). Qual papel da comunidade científica no combate à pandemia de covid-19? reflexões sobre fake news, revistas predatórias e políticas públicas.

Ferreira, A. L. N., Nascimento, D. G., Basílio, S. A. C., and SILVA, J. G. R. (2020). Um modelo matemático para classificação de fake news na web. *Anais do LII Simpósio Brasileiro de Pesquisa Operacional*.

Nathaniel, P. (2017). Can democracy survive the internet? *Journal of Democracy*, 28(2):63–76.

Oliveira, G. T. d. S. et al. (2006). Estudo e aplicações da evolução diferencial.

Storn, R. and Price, K. (1997). Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341.