Vol.xx: e00000000, 2019
http://dx.doi.org/10.1590/1678-4324-0000000000
ISSN 1678-4324 Online Edition

Artigo

Desenvolvimento de um Sistema para Classificação de Fakenews em Textos de Notícias em Língua Portuguesa

Roger Oliveira Monteiro[1]
https://orcid.org/0000-0003-1570-5794

Rodrigo Ramos Nogueira[2]
https://orcid.org/0000-0000-0000-0000

[1]Centro Universitário Leonardo da Vinci – UNIASSELVI, Capão da Canoa, Brasil; [2]University of Coimbra, Coimbra Portugal;

Associate Editor:

Received: YYYY.MM.DD; Accepted: YYYY.MM.DD.

* Correspondence: roger.o.monteiro@gmail.com; wrkrodrigo@gmail.com

Abstract: O termo fake news tem recebido preocupante atenção nos últimos anos. O crescimento das redes sociais e o fácil acesso e compartilhamento de dados na web, tem colaborado para que o problema cresça exponencialmente. Até mesmo a famosa rede social Facebook, mostrou-se inábil no combate às fake news, obtendo resultados tímidos após muitos esforços, mostrando que o problema é muito maior do que imaginávamos. Sendo assim, este trabalho tem por objetivo aplicar métodos modernos de Aprendizado de Máquina (Machine Learning) em bases textuais, como forma de buscar alternativas para criação de uma ferramenta para o combate das fake news.

Keywords: fake news; data mining; machine learning.

INTRODUÇÃO

Com o avanço da tecnologia e a explosão de dados e informações que compartilhamos e recebemos diariamente, é cada vez mais difícil um controle do que é verdadeiro e do que é falso. Segundo MONTEIRO et al. (2018), devido sua natureza atraente, as notícias falsas se espalham rapidamente, influenciando o comportamento das pessoas em diversos assuntos, desde questões saudáveis (por exemplo, revelando medicamentos milagrosos) até a política e economia, como no caso recente do escândalo Cambrige Analytica/Facebook e na situação Brexit.

Em um caso recente, a cidade de São Paulo registrou neste ano 363 casos de pessoas com sarampo, além de 800 em investigação, ou seja, um aumento de 164,9 % nas três primeiras semanas de julho, contra apenas 6% de adesão da população. O

prefeito da cidade, Bruno Covas comentou: “As pessoas postam notícias que recebem sem checar, divulgando informações sem comprovação científica.”

Segundo NOGUEIRA (2018), os sites de notícias são o terceiro maior veículo de informação mais acessado da internet, perdendo apenas para aplicativos de mensagens e redes sociais, Esta informação reflete a importância do uso de sites de notícias e se impacto no cotidiano das pessoas.

Dado a importância das notícias do cotidiano das pessoas, este trabalho tem por objetivo explorar métodos de aprendizado de máquina que visam atenuar ou até mesmo coibir a criação e disseminação de fake news através de um sistema de classificação automática.

TRABALHOS RELACIONADOS

No que se refere à notícias falsas e a aplicação de Machine Learning , GRUPPI et al. (2018) construíram um dataset com notícias, em português e inglês, tendo por objetivo construir um classificador para predizer se a fonte da notícia é ou não confiável. Utilizando um algoritmo de SVM com um kernel linear, foi possível estabelecer as características mais importantes, bem como sua classificação. Como resultado, o algoritmo de classificação obteve acurácia de 85% para os datasets brasileiros e 72% para datasets Americanos.

Em uma contribuição para a área de classificação de notícias, MONTEIRO et al. (2018) utilizam o dataset Fake.br com o objetivo de avaliar os principais métodos de pré-processamento de textos para avaliar o desempenho do método SVM. Os melhores resultados foram obtidos com a combinação de bag-of-words com sentimentos, bem como o uso de todos os atributos, ambos com acurácia de 90%.

MARUMO (2018) coletou notícias de sites com notícias verídicas e sites com notícias falsas e/ou de cunho satírico, com o objetivo de encontrar o melhor método para detecção de fake news. Como parte do pré-processamento dos dados, utilizou-se o framework Gensim para remoção de caracteres não alfabéticos, a substituição de espaçamentos e quebra de linhas para espaços únicos, remoção de palavras com menos de 3 caracteres e a conversão de letras maiúsculas para minúsculas. Também foi utilizado o framework keras para tokenização dos dados. Com a aplicação dos algoritmos de classificação LSTM e SVM, conseguiu-se uma acurácia acima de 90%.

No que se refere ao enriquecimento semântico em ambientes de Data Warehouse através do emprego de técnicas de Machine Learning , é o caso Mansman (2014), que obteve um modelo multidimensional da rede social Twitter e desenvolveu um ambiente de Data Warehouse que permitiu a criação de um cubo de dados, bem como a análise de sentimentos. Nogueira (2018), em uma abordagem similar, desenvolveu um ambiente de Data Warehouse que coleta notícias em inglês em tempo real, no qual após avaliação Regressão Logística, Naïve Bayes, SVM e Perceptron tiveram resultados próximos, dos quais o este último foi utilizado para realizar o enriquecimento semântico na etapa de ETL.

Overfitting constitui-se um grande problema em se tratando de base de dados textuais. Sendo assim, FENG, et. al. (2017), utilizaram o algoritmo *AdaBoost*, conhecido por obter grande sucesso para redução de overfitting em detecção de faces, reconhecimento de caracteres (OCR) e classificação de veículos. Em seus experimentos, foram utilizados datasets de 20 grupos de notícias, *dataset Reuters*, que consiste em 22 arquivos com um total de 21,758 documentos, e um dataset da BioMed, o qual é dividido em 10 tópicos, cada um contendo entre 1966 e 5022 artigos. Os resultados foram uma média de 86% de acurácia no algoritmo *AdaBoost* (*Bonzaiboost*).

MATERIAIS E MÉTODOS

A metodologia deste trabalho é baseada na arquitetura proposta por NOGUEIRA(2018), na qual o classificador gerado será acoplado a etapa de ETL de um Data Warehouse gerando o enriquecimento semântico em uma nova dimensão (Fig. 1).

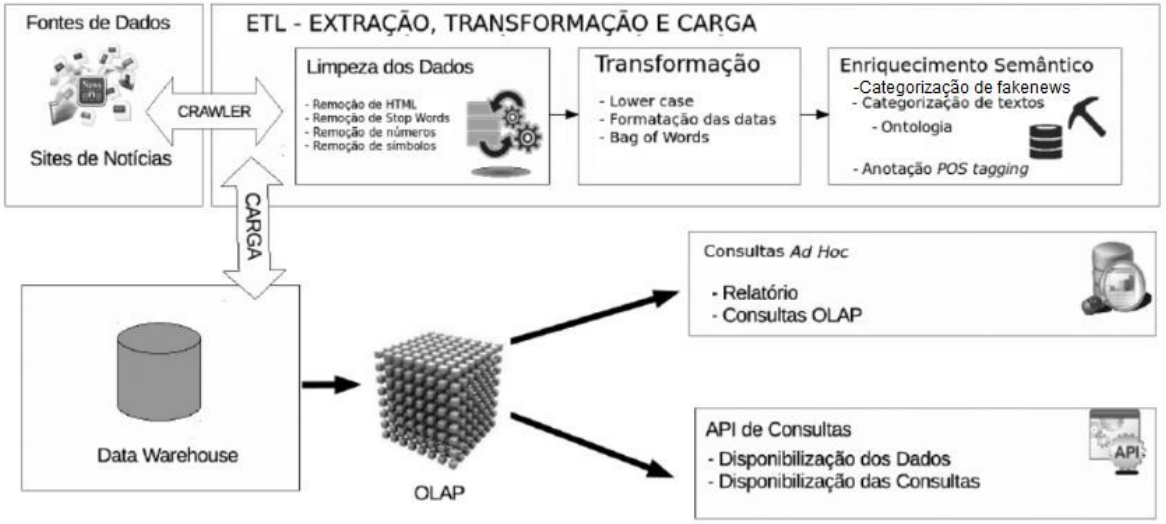

Figura 1. Arquitetura utilizada, adaptada de Nogueira (2018).

Para obtenção dos dados, fez-se a necessário uma pesquisa por sites de notícias confiáveis, e sites que veiculam ou desmentem *fake news*. Como notícias confiáveis, utilizou-se a base de dados do site *brasil.elpais.com*, e para base de dados de notícias falsas, utilizou-se o site *boatos.org* e *saude.gov.br/fakenews*. Foi então criado um *web crawler* específico para cada fonte de dados para automatizar a coleta dos mesmos. Um *web crawler* nada mais é do que um algoritmo que automatiza o processo de cópia dos dados. Foi codificado na linguagem de programação *python*, juntamente com o *framework beautiful soup*.

Os dados obtidos, contemplam o título e corpo de cada notícia. Os dados em sua forma bruta, continham ruídos, valores corrompidos, nulos, espaços em branco, caracteres especiais, etc, os quais foram tratados, limpos e transformados para letras minúsculas.

Com isso, os dados resultantes receberam uma nova coluna, chamada label, identificando quais dados são verdadeiros e quais são fake news, sendo 0 e 1 respectivamente (**Tabela 1**) e então combinados em um único dataset.

titulo	label	titulo	label
Você não consegue convencer um terraplanista e isso deveria te p	0	Vírus da febre amarela sofreu uma mutação e vacina não prc	1
DNA criado em laboratório duplica o alfabeto genético para captu	0	Correios, em 2018, contrata em site "Correios Vagas 2018" n	1
É possível recuperar o tempo de sono perdido?	0	Caseiro do sítio de Atibaia diz que Lula tem cofre de concreto	1
Por que as 'millennials' estão deixando de tomar a pílula anticonc	0	Vídeo mostra rato tomando banho com sabonete em pia	1
Os dinossauros não sumiram por conta de um meteorito	0	Lutadora de vale tudo morre após enfrentar adversária trans	1
O segredo das pessoas que perdem peso e nunca mais o recupera	0	Pirulito com energético da Blong provoca ataque cardíaco em	1
Em busca da cura para um jovem com envelhecimento precoce	0	Mais de 100 bandidos e traficantes roubaram ônibus BRT no F	1
Como a beleza afeta (injustamente) os salários e os resultados das	0	Renato Aragão, o Didi, morreu hoje de manhã	1
As incógnitas astronômicas explicadas a partir de extraterrestres	0	Macacos transmitem febre amarela para humanos e a soluçã	1
NASA dá por encerrada a missão do 'Opportunity', o robô que desl	0	Vacina contra febre amarela paralisa o fígado, diz médico de	1

Tabela 1. 10 primeiras linhas de cada dataset. A coluna label contém um valor identificador binário (0 para verdadeiro e 1 para fake news). Fonte: os autores

O dataset resultante foi dividido entre teste e treino na proporção de 75% e 25% o qual passou por um tratamento de tokenização utilizando o framework NTLK (*Natural Language ToolKit*) para a linguagem python com *bag of words* português do Brasil. Este processo faz-se necessário para que apenas palavras com significado sejam mantidas (palavras de significado sintático que não trazem informações relevantes são removidas, como por exemplo: a, ou, para, e, etc) e para que essas palavras restantes sejam classificadas conforme sentido de importância, gerando uma *parse tree ,* sendo então transformadas em números para que possam ser entendidas pelo algoritmo de *machine learning*, o qual trabalha apenas com números.

RESULTADOS

Os algoritmos de *machine learning* utilizados foram os seguintes: Regressão Logística (*Logistic Regression*), AdaBoost, Naive Bayes e SVM (*Support Vector Machines*). Todos parte integrantes do framework *sklearn*. A acurácia obtida pode ser observada na **Tabela 2**. Em bases textuais, é muito comum a ocorrência de *overfitting*, ou seja, o modelo não está de fato aprendendo com os dados, mas sim os "decorando"; durante o processo de treino atinge resultados satisfatórios, porém quando confrontado com dados reais, tem sua acurácia comprometida. Para detecção do *overfitting*, utilizou-se o método erro quadrático médio *(mean squared error)*, o qual serve para comparação de estimadores, ou seja, a diferença entre previsto e resultado.

	Regressão Logística	AdaBoost	Naive Bayes	SVM (kernel Linear)
Título	88,85%	81,37%	86,22%	87,45%
K-fold	0,88	0,75	0,86	0,55
Corpo	97,40%	95,12%	97,80%	98,62%
K-fold	0,97	0,95	0,97	0,64
Título + Corpo	90,88%	84,23%	91,19%	91,16%
K-fold	0,90	0,84	0,91	0,54

Tabela 2. Comparativo entre acurácia medida pelo algoritmo *versus* verificação com MSE.
Fonte: os autores

A partir da análise de resultados, o método de Regressão Logística utilizando o corpo das notícias foi selecionado como melhor método, pelo fato de obter uma alta acurácia, e boa tolerância ao *overfitting.*

Posterior ao acoplamento foi desenvolvido a interface de classificação de fake news, mostrada pela Figura 2. e está disponível no servidor: <https://detectorfakenews.herokuapp.com/>.

A ferramenta espera como parâmetro o link de um site de notícia, e retorna se ele é ou não uma notícia falsa (fake news).

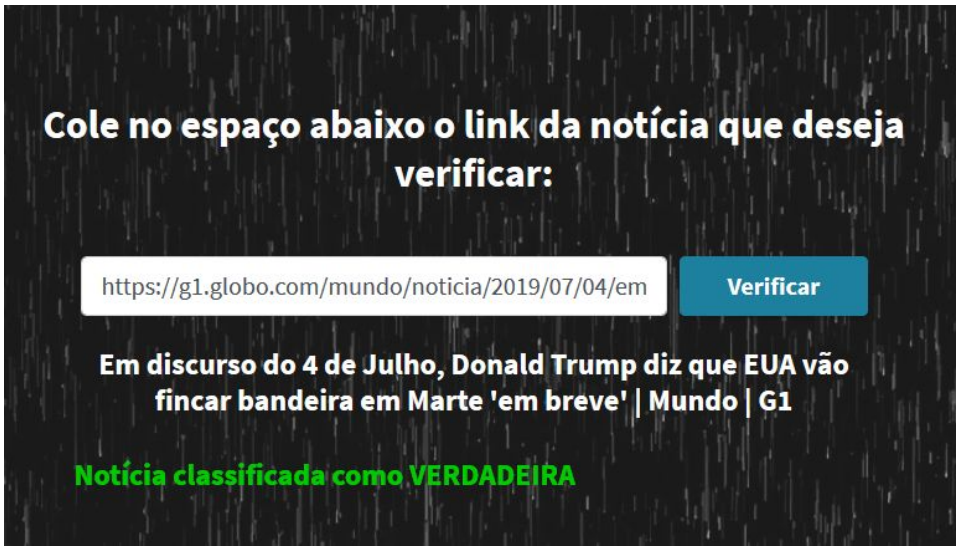

Figura 2. Interface Web da Aplicação desenvolvida. Disponível em: <https://detectorfakenews.herokuapp.com/>. Acesso em 09 julho de 2019

Os algoritmos, bem como todo restante do trabalho aqui descrito pode ser encontrado no repositório inicial <https://github.com/kerenskybr/detector_fakenews> bem como na continuidade <https://github.com/kerenskybr/detector_fakenews _2>.

DISCUSSÃO

O overfitting constitui-se um problema recorrente em bases textuais. Alguns algoritmos chegaram a resultados bastante relevantes, mas ao aplicarmos a validação cruzada com k=10, notou-se um grande overfitting em alguns casos. Sendo assim, observou-se que o algoritmo Naive Bayes obteve além da alta acurácia, tolerância ao overfitting.

Para futuros trabalhos, tem-se como objetivo avaliar outras características técnicas de pré-processamento, aumentar a base de treino, aplicar os novos resultados a interface web, e posteriormente, o acoplamento a ETL do Data Warehouse.

REFERÊNCIAS

1. DELMAZO, Caroline; VALENTE, Jonas CL. Fake news nas redes sociais online: propagação e reações à desinformação em busca de cliques. Media & Jornalismo, v. 18, n. 32, p. 155-169, 2018.
2. FENG, Xiaoyue; LIANG, Yanchun; SHI, Xiaohu; XU, Dong; WANG, Xu; GUAN, Renchu. “Overfitting Reduction of Text Classification Based on AdaBELM”, 2017
3. Fundamentals of Statistical Signal Processing: Estimation Theory by Steven M. Kay (ISBN 0-13-345711-7)
4. GRUPPI, Maurício; HORNE, Benjamin D.; ADALI, Sibel. “An Exploration of Unreliable News Classification in Brazil and The U.S.” Rensselaer Polytechnic Institute, Troy, New York, USA.2018.
5. IDC. Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyze the future, 2007(2012), 1-16.
6. <Logistic Regression: Statnotes, from North Carolina State University, Public Administration Program>. Acesso em 31 de maio de 2019.
7. MANSMANN, Svetlana; REHMAN, Nafees Ur; WEILER, Andreas; SCHOLL, Marc H. “Discovering OLAP dimensions in semi-structured data.” Information Systems, v. 44, p. 120-133, 2014.Writer’s Handbook. Mill Valley, CA: University Science, 1989.

8. MARON, M. E. (1961). "Automatic Indexing: An Experimental Inquiry" (PDF). Journal of the ACM. 8 (3): 404–417.
9. MARUMO, Fabiano Shiiti. “Deep Learning para classificação de Fake News por sumarização de texto.” - Londrina, 2018.
10. MONTEIRO, Rafael A.; SANTOS, Roney L. S.; PARDO, Thiago A. S.; ALMEIDA, Tiago A. de; RUIZ, Evandro E. S.; VALE, Oto A.. “Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results.” In: International Conference on Computational Processing of the Portuguese Language. Springer, Cham, 2018. p. 324-334.
11. NARASIMHA Murty, M.; SUSHEELA Devi, V. (2011). Pattern Recognition: An Algorithmic Approach.
12. NOGUEIRA, Rodrigo Ramos. O Poder do Data Warehouse em Aplicações ed Machine Learning: Newsminer: Um Data Warehouse Baseado em Textos de Notícias. São Paulo: Nea, 2018.
13. RUSSELL, Stuart; NORVIG, Peter (2003) [1995]. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall.
14. VAPNIK et al., 1997 e SARADHI et al., 2005).
15. VON LOCHTER, Johannes et al. Máquinas de classificação para detectar polaridade de mensagens de texto em redes sociais. 2015.