Desenvolvimento de um sistema para a classificação de Fakenews com Textos de Notícias em língua Portuguesa

Article · April 2019

CITATIONS
0

READS
332

3 authors, including:

Roger Oliveira Monteiro

4 PUBLICATIONS **0** CITATIONS

SEE PROFILE

Rodrigo Nogueira

University of Coimbra

39 PUBLICATIONS **15** CITATIONS

SEE PROFILE

Desenvolvimento de um sistema para a classificação de *Fakenews* com Textos de Notícias em língua Portuguesa

Roger Oliveira Monteiro, Rodrigo Ramos Nogueira, Greisse Moser

Centro Universitário Leonardo da Vinci – UNIASSELVI - BR 470 - Km 71

roger.o.monteiro@gmail.com,rodrigo.nogueira@uniasselvi.com.br,

greisse.moser@uniasselvi.com.br

Resumo. *Com o rápido avanço da tecnologia e o fácil acesso e disseminação de informações, o termo fakenews vem ganhando preocupante atenção e pesquisas em diversas áreas vêm sendo desenvolvidas. Sendo assim, o objetivo deste trabalho é usar métodos de aprendizado de máquina para descobrir, classificar e armazenar textos de notícias falsas, para posterior aplicação a etapa ETL de um Data Warehouse e um ambiente de consulta que contribuirá com pesquisas futuras. Para isso foi criado um dataset e os métodos Regressão Logística, Naive Bayes e SVM foram avaliados. Finalizando o trabalho com a seleção do melhor método que foi inserido em um sistema de avaliação online de notícias falsas.*

1. Introdução

Diante da facilidade com que hoje em dia qualquer pessoa pode ter acesso a informação, e com a facilidade do seu uso, vivenciamos uma era de grandes avanços e soluções, seguido porém, por problemas ainda maiores, como é o caso das notícias falsas. Segundo MONTEIRO et al. (2018), devido à 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 política e economia (como no recente escândalo Cambridge Analytica / Facebook e na situação Brexit).

Dado seu destaque, tem sido realizadas diversas multidisciplinares sobre o tema. Almejando contribuir com tais pesquisas, este trabalho tem como objetivo acoplar à etapa de ETL (*Extract, Transform, Load*) de um *Data Warehouse* de Notícias o enriquecimento semântico através de classificação do tipo de notícias: real ou falsa.

2. Trabalhos Correlatos

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. Rodando 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 falsa e/ou de cunho satírico, com o objetivo de encontrar o melhor método para detecção de fakenews. 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*.

3. Metodologia - Proposta de Aplicação

Após pesquisas por base de dados com *fakenews*, verificamos que existem poucos recursos disponíveis no idioma Português do Brasil, no qual o dataset mais utilizado é o Fake.br (MONTEIRO et al., 2018). A proposta apresentada, tem como objetivo proporcionar um ambiente com dados consistentes e limpos na forma de um corpus multidimensional para consumo por aplicações externas e usuários. O corpus multidimensional é um conjunto de textos armazenados de acordo com um modelo multidimensional, que permite explorar a multidimensionalidade em diferentes níveis de abstração: tempo, categoria das notícias, tipo (verdadeira ou *fakenews*).

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.

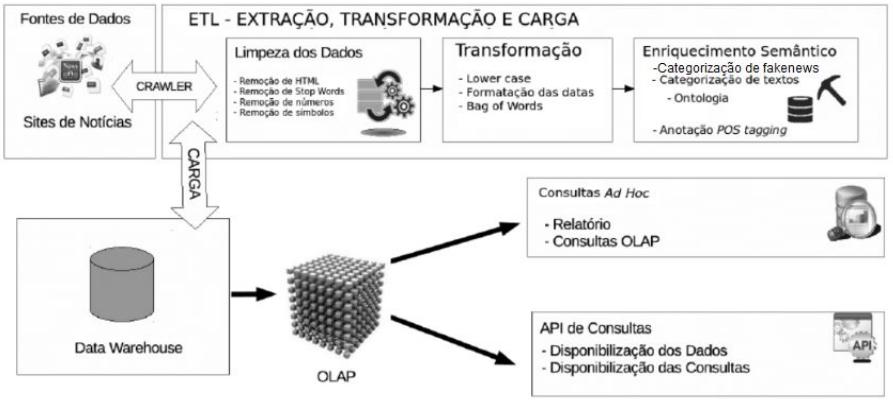

Figura 1. Arquitetura utilizada, adaptada de Nogueira (2018).

Para realizar os experimentos, foi desenvolvido um *web crawler*, utilizando a linguagem *python*, juntamente com a biblioteca *beautiful soup*, *chromium web driver* e *selenium web driver*, para a coleta inicial dos dados. Obteve-se então, um dataset composto por 2714 notícias falsas coletadas do site *boatos.org* e 3185 notícias verdadeiras coletadas do site *brasil.elpais.com*. Ambos *datasets* sendo constituídos dos títulos e corpos das respectivas notícias.

A partir da criação de um sistema de coleta, com um algoritmo acoplado à etapa de ETL, este irá automaticamente classificar os dados coletados, aumentando assim a acurácia do classificador, e gerando uma base maior de dados para futuros trabalhos de combate a *fakenews*. Também foi construído uma interface *Web*, onde o usuário será capaz de submeter um *link* e verificar se este é ou não uma notícia verdadeira, servindo este como protótipo antes de ser submetido a etapa de *ETL*(sendo esta, o propósito geral deste trabalho).

	titulo	corpo
0	Você não consegue convencer um terraplanista e...	Terraplanistas fretam cruzeiro para ir até a b...
1	DNA criado em laboratório duplica o alfabeto g...	Toda a informação necessária para montar e ope...
2	A Microsoft encerra a era Bill Gates	Já é oficial. Na manhã desta terça-feira em Se...
3	Facebook dispara seu lucro líquido para cerca ...	O lucro líquido do Facebook chegou a cerca de ...
4	Uma líder discreta	Michelle Bachelet, pediatra, de anos, volta a...
5	Só consideram provocação quando desenhamos Maomé	Leia abaixo reportagem sobre o diretor da "Cha...

	titulo	corpo
0	Sérgio Moro vai sair do governo, não esta co...	BOMBÁSTICO: SERGIO MORO VAI SAIR DO ...
1	O homem que agrediu Manu, era segurança de Bo...	Que zorra é essa ? Não é aquele bolsonaris...
2	"A Atriz Adriana Esteves sendo atacado por um ...	"A Atriz Adriana Esteves foi atacada por um C...
3	Após ser roubada, mulher obriga ladrão a man...	uma mulher identificada como Mônica Santos, c
4	Brasileiro leva surra ao tentar gravar vídeo ...	Esse é o Fábio, um típico Playboy Filhinho ...
5	Jacaré gigante é assassinado no Marajó	Gigante jacaré Açu foi visto e morto por pes...

Tabela 1. Cinco primeiras linhas de ambos datasets.

Posteriormente, utilizando a literatura de referência, foram selecionados quatro métodos para serem avaliados no dataset: *Gradient Boosting, Logistic Regression, Naive Bayes e SVM*. O método Gradient Boosting foi incluído unicamente por se tratar de um método *Ensemble,* ou seja, uma mistura de classificadores múltiplos, comumente utilizado em bases propensas a *overfitting*. Após a avaliação o melhor método será acoplado à etapa de ETL do sistema proposto, bem como a interface *Web* de classificação de notícias.

4. Resultados Parciais

Inicialmente, o experimento contou apenas com a utilização dos títulos das notícias. Os dados então receberam tratamento de valores nulos, ruídos (caracteres especiais, tais como vírgulas, pontos, parênteses, resíduos de *tags html*, etc) e transformação para letras minúsculas. Cada dataset recebeu uma nova coluna, chamada label, onde foi atribuído o valor *booleano* 0 para notícias verdadeiras, e 1 para as notícias falsas, sendo esta coluna a variável preditora, que irá classificar os novos dados de acordo. Com isso,

os dados foram combinados em um único dataset.

O *dataset* foi então dividido entre treino e teste, nas proporções de 85% e 15%, 75% e 25%, e 67% e 33% respectivamente para que testes sejam realizados para estabelecer a melhor proporção a ser adotada. A primeira parcela dos dados serve para treinar o algoritmo, enquanto a segunda, como testes para verificar a acurácia do mesmo. Na sequência, os dados receberam tratamento de tokenização, utilizando o pacote *NLTK*, com o *bag of words* em português do Brasil, que transforma palavras em valores numéricos, pois um algoritmo consegue lidar apenas com números.

Testes efetuados utilizando apenas os títulos das notícias com os algoritmos *Gradient Boosting*, Regressão Logística (*Logistic Regression*), *Naive Bayes* e *SVM* (*kernel linear*), obtiveram a acurácia conforme **Tabela 2**, no modelo de testes.

Os resultados parciais obtidos após a construção, treino e produção do modelo foram satisfatórios. O algoritmo escolhido para a implementação inicial foi o *SVM*, que além de obter o melhor desempenho, mostrou-se bastante recorrente na literatura consultada. Como técnica de avaliação do modelo empregado, foi utilizado a validação cruzada com o método *k-fold* = 10.

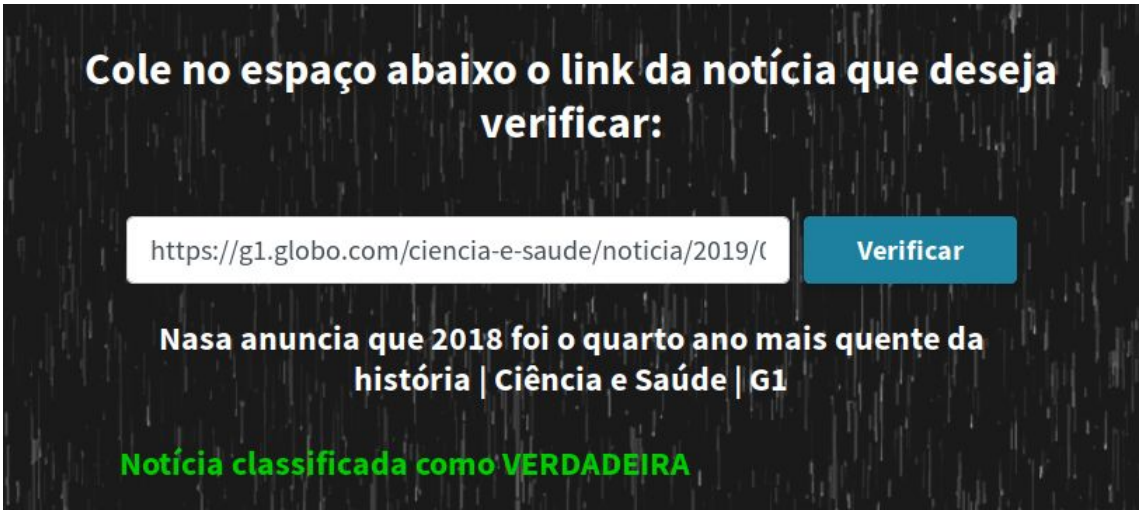

Figura 2. Interface Web da Aplicação desenvolvida. Disponível em: <https://detectorfakenews.herokuapp.com/>. Acesso em 18 fev. 2019.

Para utilização do corpo da notícia, novamente os dados passaram pelo mesmo tratamento empregado aos títulos. Um problema muito comum quando utilizamos bases de dados textuais, é o *overfitting,* ou seja, o algoritmo funciona muito bem quando testamos com o conjunto de testes, mas mostra-se ineficaz quando aplicado a dados reais. Como forma de verificar onde o problema ocorre, podemos aplicar o *MSE* (*Mean Square Error*), para verificar a margem de erro entre o que deveria ser e o que foi predito.

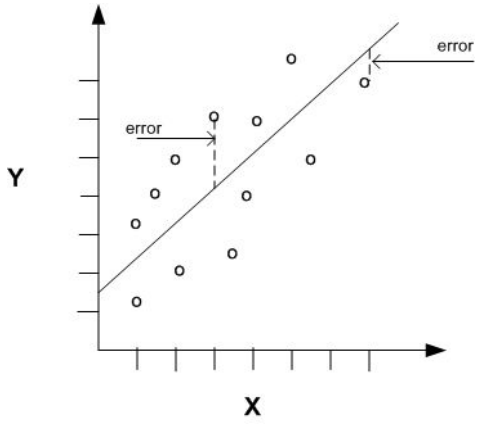

Figura 3. MSE (*Mean Square Error)*. Disponível em: <https://educationalresearchtechniques.com/2015/11/>

Acesso em 22 mar. 2019.

Um comparativo de treinos e testes foi efetuado para determinar o melhor método a ser utilizado. Os resultados podem ser observados na **Tabela 2**.

Testes utilizando apenas o título da notícia								
Treino / Teste 85 / 15	Acurácia	MSE	Treino / Teste 75 / 25	Acurácia	MSE	Treino / Teste 67 / 33	Acurácia	MSE
Gradient Boosting	61.78	0.3821	Gradient Boosting	59.94	0.4005	Gradient Boosting	60.07	0.3992
SVM (Kernel Linear)	91.57	0.0842	SVM (Kernel Linear)	91.15	0.0884	SVM (Kernel Linear)	90.51	0.0948
Logistic Regression	90.31	0.0968	Logistic Regression	90.96	0.0903	Logistic Regression	90.18	0.0981
Naive Bayes	89.78	0.1021	Naive Bayes	89.82	0.1017	Naive Bayes	88.89	0.1111

Testes utilizando o corpo notícia								
Treino / Teste 85 / 15	Acurácia	MSE	Treino / Teste 75 / 25	Acurácia	MSE	Treino / Teste 67 / 33	Acurácia	MSE
Gradient Boosting	77,48	0,2252	Gradient Boosting	81,80	0,1829	Gradient Boosting	79,38	0,2062
SVM (Kernel Linear)	96,85	0,0315	SVM (Kernel Linear)	96,94	0,0306	SVM (Kernel Linear)	97,46	0,0254
Logistic Regression	97,57	0,0242	Logistic Regression	97,23	0,0277	Logistic Regression	97,46	0,0254
Naive Bayes	98,78	0,0121	Naive Bayes	97,81	0,0218	Naive Bayes	98,23	0,0176

Tabela 2. Testes mostrando a porcentagem de treino/teste, a acurácia e o MSE.

5. Considerações Finais e Trabalhos Futuros

Os testes revelaram que bases textuais são muito propensas ao *overfitting*. Existe a necessidade de encontrar melhores métodos para amenizar este problema em trabalhos

futuros. O algoritmo *Naive Bayes* mostrou maior acurácia e menor margem de erro quando aplicado a dados reais. Nos testes efetuados, o algoritmo acertou todas as vezes. A utilização do corpo das notícias mostrou-se mais eficaz quando comparado a utilização apenas do título, porém ao custo de maior tempo de processamento.

O estudo mostrou-se relevante para o aperfeiçoamento e entendimento dos envolvidos, bem como a corroboração da necessidade do combate às *fake news.* 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,* bem como ampliar a compatibilidade da mesma com os mais diversos sites, e posteriormente, o acoplamento a ETL do *Data Warehouse.*

Referências

KURACH, Karol & Pawłowski, Krzysztof & Romaszko, Łukasz & Tatjewski, Marcin & Janusz, Andrzej & Nguyen, Hung Son. (2012). An Ensemble Approach to Multi-label Classification of Textual Data. 7713. 306-317. 10.1007/978-3-642-35527-1_26.

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.

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.

MARUMO, Fabiano Shiiti. “Deep Learning para classificação de Fake News por sumarização de texto.” - Londrina, 2018.

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.

NOGUEIRA, Rodrigo Ramos. (2017). Newsminer: um sistema de data warehouse baseado em textos de notícias. Universidade Federal de São Carlos (UFSCar)