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.

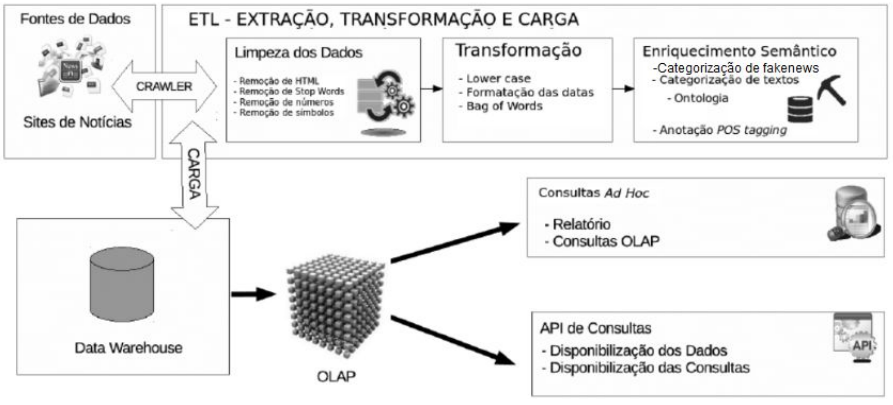

Figura 1. Arquitetura utilizada, adaptada de Nogueira (2018).

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. Foi construído um dataset composto por 2714 títulos de notícias falsas coletadas do site *boatos.org* e 3185 títulos de notícias verdadeiras coletadas do site *brasil.elpais.com*. Inicialmente será utilizado apenas os títulos das notícias. Posteriormente, planejamos a utilização da notícia por inteiro.

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_noticia	url	label
0	Indústria brasileira rea ...	https://brasil.elpais.com/brasil/20:...	0
1	A bancarrota de Detro:...	https://brasil.elpais.com/brasil/20:...	0
2	PIB no Brasil cai 0...	https://brasil.elpais.com/brasil/20:...	0
3	O órgão supervisor eur:...	https://brasil.elpais.com/brasil/20:...	0
4	Vega S...	https://brasil.elpais.com/brasil/20:...	0

	titulo_noticia	url	label
0	Neto de Chico Buarque f...	www.boatos.org/entretenim|...	1
1	Correios, em 2018, c...	www.boatos.org/tecnologia...	1
2	Caseiro do sítio de ,...	www.boatos.org/politic-...	1
3	Vídeo mostra rato tomanc...	www.boatos.org/mundo/video....	1
4	Lutadora de vale tudo:...	www.boatos.org/esporte/lutr...	1

Tabela 1. Cinco primeiras linhas de ambos datasets.

Posteriormente, utilizando a literatura como referência foram selecionados três métodos para serem avaliados no dataset: Regressão Logística (Logistic Regression), Naive Bayes e SVM. 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

Os dados obtidos receberam tratamento de valores nulos, ruídos (caracteres especiais, tais como vírgulas, pontos, parênteses, 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. Com isso, os dados foram combinados em um único dataset. Os rótulos das colunas foram convertidos em valores numéricos utilizando o Label Encoder do pacote scikit-learn.

O dataset foi então dividido entre treino e teste, na proporção de 75% e 25% respectivamente. A primeira parcela serve para treinar o algoritmo, enquanto a segunda, para verificar a acurácia do mesmo. Na sequência, receberam tratamento de tokenização, utilizando o pacote NLTK, com o *bag of words* em português do Brasil.

Testes efetuados utilizando os algoritmos Regressão Logística (Logistic Regression), Naive Bayes e SVM (kernel linear), obtiveram a acurácia de 90.33%, 89.27% e 90.52% respectivamente, 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.

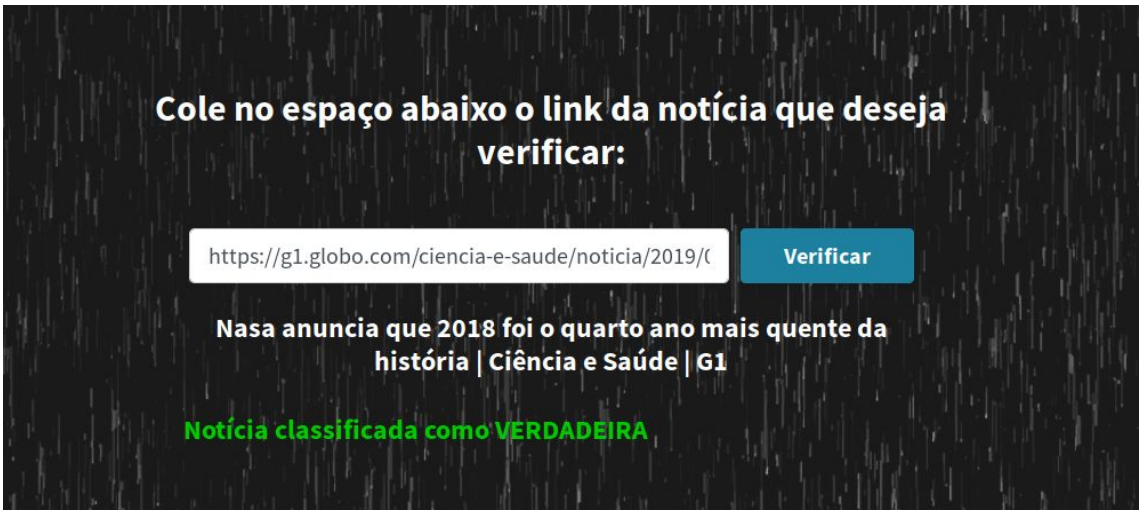

Figura 2. Interface Web da Aplicação desenvolvida. Disponível em: <https://detectorfakenews.herokuapp.com/>. Acesso em 18 fev. 2018.

5. Considerações Finais e Trabalhos Futuros

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, utilizar além do título, a notícia por completo, aplicar os novos resultados a interface *web,* e posteriormente, o acoplamento a ETL do *Data Warehouse.*

Referências

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. O Poder do Data Warehouse em Aplicações ed Machine Learning: Newsminer: Um Data Warehouse Baseado em Textos de Notícias. São Paulo: Nea, 2018.