Projeto de um Sistema Web a Classificação de Fake News

Roger Oliveira Monteiro
Centro Universitário Leonardo da Vinci
Osório, Brasil
e-mail: roger.o.monteiro@gmail.com

Rodrigo Ramos Nogueira
University of Coimbra
Coimbra, Portugal
e-mail: rodrigonogueira@dei.uc.pt

***Abstract* – A internet soma mais de 2 bilhões de sites publicados, sendo a principal fonte de informação deste século. No entanto, cada vez mais sites implicam em diversos veículos que não produzem notícias verdadeiras, mas sim falsas, as ditas fakes news. Tendo em vista realizar a classificação automática de fake news este artigo apresenta um sistema que realiza a coleta e classificação de notícias. Para isto, utiliza 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. Por fim, o melhor algoritmo foi acoplado a um sistema web que realiza a classificação de fake news baseado em aprendizado automático.**

Keywords: fake news, machine learning, data warehouse.

I. INTRODUÇÃO

Desde o início da Web, o volume de dados que estão nos repositórios na rede mundial tem crescido de forma exponencial. Atualmente, são cerca de 200 milhões de sites ativos na Internet, dos quais, apenas a rede social Twitter gera em média, 500 milhões de postagens por dia. Tal explosão de dados, levou a um estudo do IDC (Institute Data Corporation) que estima que até 2020 serão gerados 44 zettabytes de dados em todo mundo (IDC, 2012).

Nos diferentes nichos de redes sociais que surgiram, observou-se maneiras diferentes de redigir críticas, propiciadas pelas características das aplicações. Sites específicos, como especializados em críticas de filmes, permitem que usuários escrevam textos relativamente longos. Os microblogs, por outro lado, impõem limites na quantidade de caracteres das mensagens e não são ambientes exclusivamente destinados para publicação de críticas. No processo de descoberta e pesquisa que prosseguiu nas redes sociais, surgiu a necessidade de expressar opiniões de forma mais direta (VON LOCHTER, 2015).

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 seu impacto no cotidiano das pessoas.

Associado a importância de textos de notícias e seu compartilhamento das mesmas em redes sociais, vem a ascensão e disseminação das fake news. Desde meados de 2017, a quantidade de eventos e debates acerca deste fenômeno que vem sendo chamado de *fake news* tornou-se recorrente. Fake news pode ser definida como artigos de notícias que são intencional e verificadamente falsos e podem enganar os leitores. E essa definição de fake news inclui artigos de notícias fabricados intencionalmente, como um artigo amplamente compartilhado do agora extinto site *denverguardian.com* com a manchete "*FBI agent suspected in Hillary email leaks found dead in apparent murder-suicide*" (Agente do FBI suspeito de vazamento de e-mail de Hillary encontrado morto em aparente assassinato-suicídio) (DELMAZO, 2017).

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.

II. 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. 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*).

III. DESENVOLVIMENTO

Após pesquisas por base de dados com *fake news*, 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 fake news).

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).

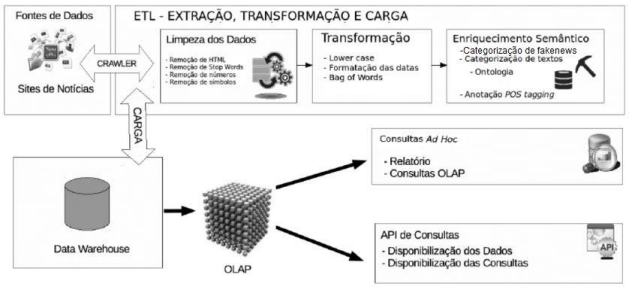

Figure 1. Arquitetura utilizada, adaptada de Nogueira (2018).

Para realizar os experimentos foi desenvolvido um *web crawler*, utilizando a linguagem *python*, juntamente com a biblioteca *beautiful soup* para a coleta inicial dos dados. Como resultado, obteve-se um *dataset* composto por 1744 títulos e corpos de notícias falsas coletadas dos sites *boatos.org* e *g1.globo.com/fato-ou-fake*, e 3185 títulos e corpo de notícias verdadeiras coletadas do site *brasil.elpais.com*. Testes serão efetuados utilizando apenas os títulos das notícias, o corpo, e então uma junção de ambos, fazendo assim um comparativo entre estes. Para isso, serão utilizados os algoritmos de aprendizado de máquina (*Machine Learning*), Regressão Logística, AdaBoost, Naive Bayes e SVM.

A Regressão Logística (*Logistic Regression*) é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias. É útil para modelar a probabilidade de um evento ocorrer como função de outros factores. A regressão logística analisa dados distribuídos binomialmente da forma

$$Yi \sim B(pi, ni), for\ i = 1, \ldots, m, \quad (1)$$

onde os números de ensaios de Bernoulli *ni* são conhecidos e as probabilidades de êxito *pi* são desconhecidas. Um exemplo desta distribuição é a percentagem de sementes (*pi*) que germinam depois de *ni* serem plantadas.

O *AdaBoost* é um algoritmo meta-heurístico, e pode ser utilizado para aumentar a performance de outros algoritmos de aprendizagem. O *AdaBoost* chama um classificador fraco repetidamente em iterações

$$t = 1, \ldots, T. \quad (2)$$

Para cada chamada a distribuição de pesos *Dt* é atualizada para indicar a importância do exemplo no conjunto de dados usado para classificação. A cada iteração os pesos de cada exemplo classificado incorretamente é aumentado (ou alternativamente, os pesos classificados corretamente são decrescentes), para que então o novo classificador trabalhe em mais exemplos.

No aprendizado de máquina, classificadores *Naive Bayes* são uma família de simples "classificadores probabilísticos" baseados na aplicação do teorema de Bayes com pressupostos de independência fortes (naive) entre as características. Os classificadores *Naive Bayes* são altamente escalonáveis, exigindo um número de parâmetros lineares no número de variáveis (recursos/preditores) em um problema de aprendizado. O treinamento de máxima verossimilhança pode ser feito através da avaliação de uma expressão de forma fechada, que leva um tempo linear, em vez de uma aproximação iterativa dispendiosa como usada para muitos outros tipos de classificadores. Abstratamente, *Naive Bayes* é um modelo de probabilidade condicional: dada uma instância de problema a ser classificada, representada por um vetor

$$x = (x1, \ldots, xn) \quad (3)$$

representando alguns n recursos (variáveis independentes), atribui a esta instância probabilidades

$$p(Ck \mid x1, \ldots, xn) \quad (4)$$

para cada um dos K possíveis resultados ou classes *Ck* .

O *Support Vector Machine* (SVM), também conhecido como Máquina de Suporte Vetorial, foi elaborado com o estudo proposto por Boser, Guyon e Vapnik em 1992. É um algoritmo de aprendizado supervisionado, cujo objetivo é classificar determinado conjunto de pontos de dados que são mapeados para um espaço de características multidimensional usando uma função *kernel*, abordagem utilizada para classificar problemas. Nela, o limite de decisão no espaço de entrada é representado por um hiperplano em dimensão superior do espaço. No caso do *kernel* linear, recebemos um conjunto de dados de treinamento de n pontos da forma

$$(x1,y1),\ldots,(xn,yn), \quad (5)$$

onde os *yi* são 1 ou -1, cada um indicando a classe à qual o ponto *xi* pertence. Cada *xi* é um *p*- vetor real tridimensional. Queremos encontrar o "hiperplano de margem máxima" que divide o grupo de pontos xi para qual *yi = 1* do grupo de pontos para os quais *yi = -1*, que é definido de modo que a distância entre o hiperplano e o ponto mais próximo *yi* de qualquer um dos grupos é maximizado.

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).

IV. Resultados Parciais

Inicialmente, o *dataset* utilizado continha apenas os títulos das notícias, sendo 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*), *AdaBoost*, *Naive Bayes* e SVM (kernel linear), obtiveram a acurácia de 88.85%, 81.37%, 86.22% e 87.45% respectivamente, no modelo de testes. Como técnica de avaliação dos modelos empregados, foi utilizado a validação cruzada com o método k-fold = 10 (Fig. 2).

Novamente o *dataset* foi dividido entre treino e teste, juntando agora os títulos ao corpo das notícias. Receberam os mesmos tratamento acima citados, obtendo a acurácia de 90.88%, 84.23%, 91.19% e 91.16% nos algoritmos Regressão Logística (*Logistic Regression*), *AdaBoost*, *Naive Bayes* e *SVM* respectivamente. A aplicação do método de validação cruzada, revelou um *overfitting* em alguns casos.

Por fim, o *dataset* foi dividido para utilização apenas dos corpos das notícias. Foram empregados os mesmos métodos utilizados anteriormente em relação ao tratamento e limpeza dos dados. A aplicação dos algoritmos resultou em 90.88%, 94.23%, 91.19% e 91.16% de acurácia nos algoritmos Regressão Logística (*Logistic Regression*), *AdaBoost*, *Naive Bayes* e SVM respectivamente.

	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

Figure 2. Comparativo entre os datasets em relação à acurácia e método de validação cruzada.

A partir da análise de resultados, o método de Naive Bayes foi selecionado o melhor método, pelo fato de obter uma alta acurácia, complementado de ser um método de aprendizado incremental (online).

Posterior ao acoplamento foi desenvolvido a interface de classificação de *fake news*, mostrada pela Fig 3. 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).

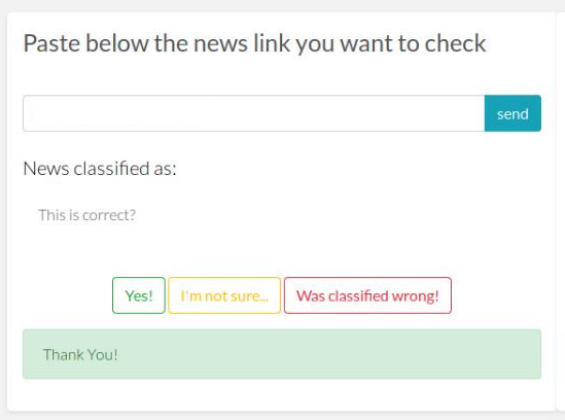

Figure 3. Interface Web da Aplicação desenvolvida. Disponível em: <https://detectorfakenews.herokuapp.com/>. Acesso em 09 julho de 2019

V. Considerações Finais e Tarbalhos Futuros

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*.

REFERENCIAS

[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] 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.

[4] 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.

[5] <Logistic Regression: Statnotes, from North Carolina State University, Public Administration Program>. Acesso em 31 de maio de 2019.

[6] 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.

[7] MARON, M. E. (1961). "Automatic Indexing: An Experimental Inquiry" (PDF). Journal of the ACM. 8 (3): 404–417.

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

[9] 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.

[10] NARASIMHA Murty, M.; SUSHEELA Devi, V. (2011). Pattern Recognition: An Algorithmic Approach.

[11] 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.

[12] RUSSELL, Stuart; NORVIG, Peter (2003) [1995]. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall.

[13] VAPNIK et al., 1997 e SARADHI et al., 2005).

[14] VON LOCHTER, Johannes et al. Máquinas de classificação para detectar polaridade de mensagens de texto em redes sociais. 2015.