Detecção e Classificação de Fake News

Isabella Andrade, Heriberto Alexandre

Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco (UFRPE) - 52171-900 – Recife – PE – Brasil

isabella.fernandes@ufrpe.br
heriberto.alexandre@ufrpe.br

Abstract. *This article aims to study the classification of false news through the use of artificial intelligence algorithms and datasets available on the Internet. There are also comparisons made between the results obtained by these algorithms.*

Resumo. *Este artigo tem como objetivo estudar a classificação de notícias falsas através do uso de algoritmos de Inteligência Artificial e de datasets disponíveis na internet. Também são realizadas comparações entre os resultados obtidos por esses algoritmos.*

Keywords. *Inteligência Artificial, Machine Learning, Fake News, Naive Bayes, Random Forest.*

1. Introdução

As chamadas Notícias Falsas (ou Fake News) são notícias escritas e publicadas com a intenção de enganar, a fim de se obter ganhos financeiros ou políticos, muitas vezes com manchetes sensacionalistas, exageradas ou evidentemente falsas para chamar a atenção (SCHLESINGER, 2017).

Seu uso têm crescido nos últimos anos, principalmente durante períodos em que a manipulação de notícias pode ter efeitos na sociedade. Por exemplo, de 2015 para 2016, ano de eleição dos Estados Unidos, enquanto a média do consumo de notícias sérias aumentou de 5,7 artigos por dia para 8,1 por dia, a média do consumo de Fake News aumentou de 8 artigos por dia para 18,3 por dia (GUESS, 2018).

Ao mesmo tempo em que a exposição a Fake News aumenta, durante quatro décadas foram realizadas pesquisas sobre detecção de fraudes para contabilizar o quão bem os humanos são capazes de detectar mentiras no texto. As descobertas mostram que não somos tão bons nisso. De fato, apenas 4% melhor que o acaso, com base em uma meta-análise de mais de 200 experimentos (BOND, 2006).

Esse problema levou pesquisadores e desenvolvedores técnicos a olhar para várias maneiras automatizadas de avaliar o valor de verdade de um texto potencialmente enganoso baseado nas propriedades do conteúdo e nos padrões de comunicação mediada por computador (CONROY, 2015). Este artigo tem como objetivo explorar a automatização de detecção de Fake News através de algoritmos de Inteligência Artificial.

2. Materiais

Neste trabalho foram utilizados dois datasets, um deles para a detecção de fake news e outro para a classificação. Em ambos os casos dividimos o dataset na proporção 80/20 para obtermos dados para treino e para teste respectivamente.

O primeiro dataset é o Fake News, utilizado para a detecção de notícias falsas, e pode ser encontrado no site https://www.kaggle.com/c/fake-news/overview. Este dataset possui 20.800 linhas e 4 colunas que estão listadas logo abaixo.

Coluna	Descrição
Id	identificação da linha
Title	título da notícia
Author	autor da notícia
Text	conteúdo da notícia

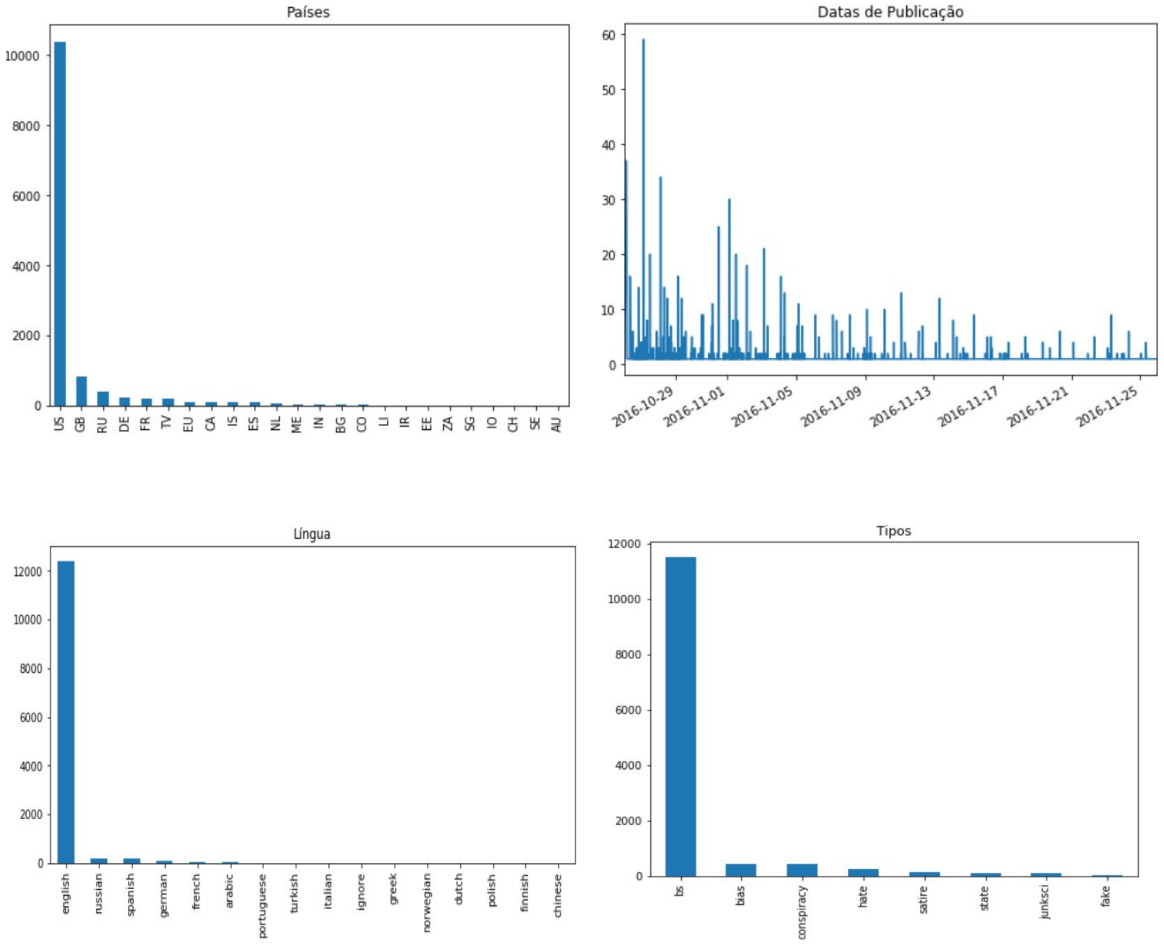
O segundo dataset é o Getting Real about Fake News, utilizado para classificação, e pode ser encontrado no site https://www.kaggle.com/mrisdal/fake-news. Este dataset possui 12999 linhas e 20 colunas, as principais estão listadas abaixo.

Coluna	Descrição
uuid	identificação da linha
author	autor da notícia
published	data em que a notícia foi publicada
title	título da notícia
text	conteúdo da notícia
site_url	link do site da notícia
country	país que publicou a notícia
type	classificação da notícia

Além disso, há 7 categorias de classificação de notícias.

Tipo	Descrição
Fake News (Fake)	Fontes que fabricam notícias de todo os tipos com a intenção de enganar o público.
Sátira (Satire)	Fontes que fornecem comentários humorísticos sobre eventos atuais na forma de notícias falsas.
Viés extremo (Bias)	Fontes que trafegam propaganda política e distorções grosseiras de fatos.
Teoria da conspiração (Conspiracy)	Fontes que são bem conhecidas por promoverem teorias conspiratórias.
Notícias do Estado (State)	Fontes em estados repressivos que operam sob sanção do governo.
Grupo de Ódio (Hate)	Fontes que promovem ativamente o racismo, a misoginia, a homofobia e outras formas de discriminação.
Ciência Lixo (Junk Science)	Fontes que promovem pseudociência, metafísica, falácias naturalistas e outras afirmações cientificamente duvidosas.

A partir de gráficos gerados com os dados do dataset Getting Real about Fake News, é possível perceber que a maior parte das notícias é em inglês e possui origem nos Estados Unidos, no ano em que ocorreram eleições presidenciais (2016).



3. Métodos

3.1 Frequência das Palavras

Para classificar as notícias como falsas ou reais, utilizamos a coluna Text do dataset, transformando os textos em vetores e contabilizando as frequências das palavras. Para essa contabilização, foram utilizados os seguintes métodos:

3.1.1 Bow

O modelo bag-of-words é um dos mais simples e de mais fácil entendimento. Neste modelo, um texto (como uma frase ou um documento) é representado como o saco (multiset) de suas palavras, desconsiderando a gramática e mesmo a ordem das palavras, mas mantendo a multiplicidade.

3.1.2 Term Frequency (TF)

Frequência de termos (TF) frequentemente usada em Mineração de Texto, PNL e Recuperação de Informações informa com que frequência um termo ocorre em um

documento. No contexto da linguagem natural, os termos correspondem a palavras ou frases. Como cada documento é diferente em tamanho, é possível que um termo apareça com mais frequência em documentos mais longos do que em documentos mais curtos.

3.1.3 TF-IDF

TF-IDF, abreviação do termo frequência do termo - inversão de frequência do documento, trata-se de uma estatística numérica cujo o objetivo é refletir a importância de um palavra no texto onde a mesma encontra-se. Assim como em TF, o número de vezes que uma palavra aparece no documento é utilizado como parâmetro para classificar sua importância, no entanto, TF-IDF aumenta proporcionalmente ao número de vezes que uma palavra aparece no documento e o valor também é compensada pelo número de documentos no escopo que contém a palavra, o que ajuda a ajustar o fato de que algumas palavras aparecem com mais frequência em geral.

3.2 Algoritmos de Classificação

Para cada método utilizado para medir a frequência das palavras, utilizamos dois algoritmos para classificar os dados.

3.2.1 Naive Bayes

O Naive Bayes é um classificador de aprendizagem de máquina simples, mas eficaz e comumente usado. Baseado na regra de Bayes, a qual é associada a suposição de independência condicional, o modelo pode ser usado com eficácia em classificação de documentos de texto, diagnóstico entre outros.

3.2.2 Random Forest

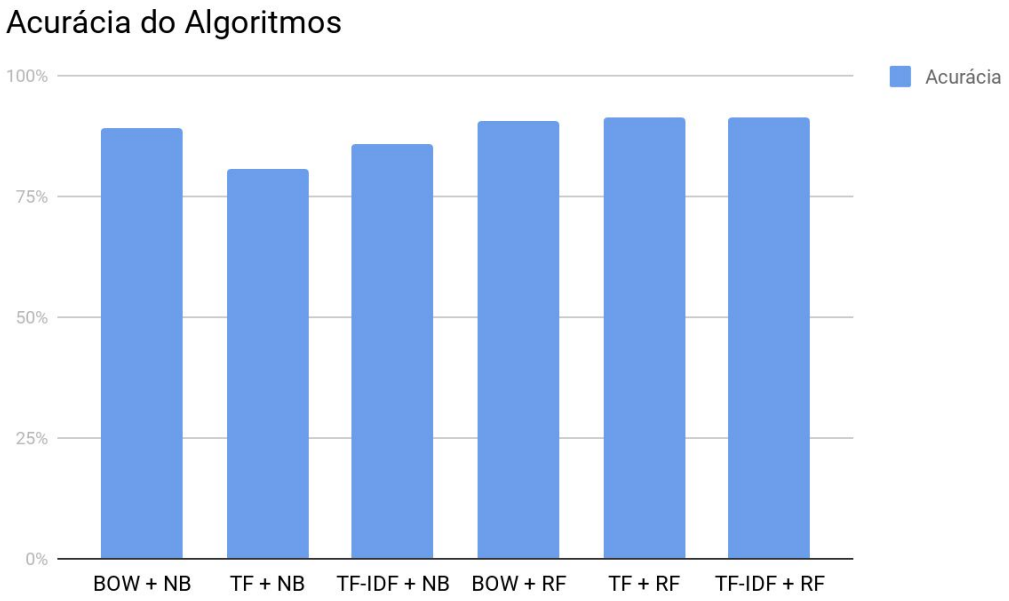
Para entender este algoritmo, primeiramente é necessário entender o conceito de uma árvore de decisão. Uma árvore de decisão pode ser compreendida como uma sequência de perguntas de sim e não, em que cada resposta leva a um nó.

Cada nó possui uma classificação definida de acordo com a quantidade de exemplos de cada classe, tal que a classe que possuir mais exemplos é a escolhida. Além da classificação, os nós possuem um índice de impureza, que representa a chance de um elemento ser classificado incorretamente. Quanto mais profundo o nível da árvore, menor o índice, até que chegue a 0 no último nível.

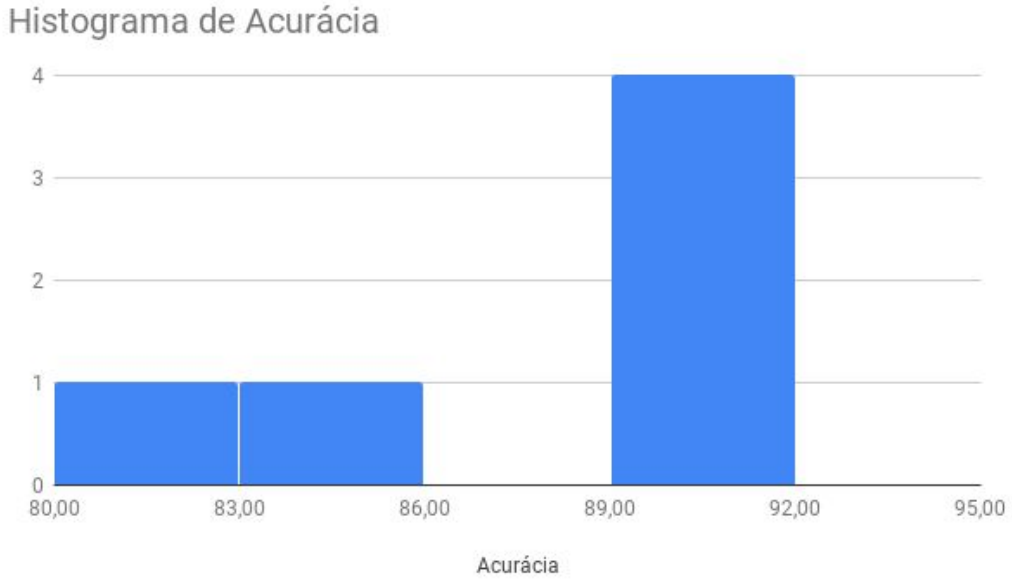
O algoritmo Random Forest utiliza várias árvores de decisão. No momento de treino, são criadas várias árvores a partir de amostras aleatórias dos dados, e para realizar as predições, o algoritmo considera todas as árvores e calcula a média para definir a classe final.

4. Experimentos e Resultados

Para avaliar as combinações dos métodos, foram verificadas as suas acurácias através da funcionalidade da própria biblioteca utilizada. O resultado entre os algoritmos foi semelhante, sendo o menor o resultado do algoritmo Term Frequency + Naive Bayes, que correspondeu a 80,55% de acurácia, e o maior resultado do algortimo TF-IDF + Random Forest, que correspondeu a 91,22%



Ainda, constatamos que a maior parte dos resultados, quatro das seis combinações, se encontra entre o intervalo de 89% a 92% de acurácia.



5. Conclusão

Após a realização de diversos testes de modelos de contagem e algoritmos combinados ficou comprovado que apesar da sua simplicidade e de sua fácil implementação o modelo bag-of-words (BOW) mostrou possuir uma eficiência mais estável comparado com outros modelos, pois sua acurácia teve grande vantagem no algoritmo Naive Bayes, e no Random Forest apesar de ter sido o mais baixo, não foi por muito. Quanto aos algoritmos de classificação, o Random Forest obteve melhor desempenho frente ao Naive Bayes independentemente do modelo de contagem.

Referências

SCHLESINGER, Robert; Fake News in Realit. US News, 2007. Disponível em <https://www.usnews.com/opinion/thomas-jefferson-street/articles/2017-04-14/what-is-fake-news-maybe-not-what-you-think>. Acesso em 06 de julho de 2019.

GUESS, Andrew; NYHAN, Brendan; REIFLER, Jason, Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign. European Research Council 9, 2018. Disponível em <http://www.ask-force.org/web/Fundamentalists/Guess-Selective-Exposure-to-Misinformation-Evidence-Presidential-Campaign-2018.pdf>. Acesso em 07/07/2019.

BOND, C. F.; DEPAULO, B. M., Accuracy of Deception Judgments. Personality and Social Psychology Review, vol. 10, 2006, pp. 214-234.

CONROY, N. J; RUBIN, V. L.; CHEN, Y., Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, vol. 52, 2015, pp. 1-4.