Classificação de *fake news* utilizando Aprendizado de Máquina

Murilo Luis C. Neves[1], **Vitor Padovani**[1], **Fernando Silva Grande**[1]

[1]Departamento de Informática – Universidade Estadual de Maringá

Abstract. *This meta-paper describes the style to be used in articles and short papers for SBC conferences. For papers in English, you should add just an abstract while for the papers in Portuguese, we also ask for an abstract in Portuguese ("resumo"). In both cases, abstracts should not have more than 10 lines and must be in the first page of the paper.*

Resumo. *Este meta-artigo descreve o estilo a ser usado na confecção de artigos e resumos de artigos para publicação nos anais das conferências organizadas pela SBC. É solicitada a escrita de resumo e abstract apenas para os artigos escritos em português. Artigos em inglês deverão apresentar apenas abstract. Nos dois casos, o autor deve tomar cuidado para que o resumo (e o abstract) não ultrapassem 10 linhas cada, sendo que ambos devem estar na primeira página do artigo.*

1. Introdução

Falar um pouco sobre fake news e aprendizado de máquina...

2. Fundamentação Teórica

2.0.1. Random Forest

2.0.2. Logistic Regression

2.0.3. Gradient Boosting

2.0.4. Redes Neurais

3. Materiais e métodos

Nesta seção, são descritos os materiais utilizados para o trabalho (seção 3.1) e os métodos como foram trabalhados (seção 3.2).

3.1. Materiais

Os materiais utilizados para esse trabalho incluem o *dataset* WELFake, e, como ferramentas de codificação principais, as bibliotecas *Scikit-Learn* e *Torch*.

3.1.1. Base de dados

O *dataset* utilizado é o WELFake, devido ao grande volume de dados presentes no mesmo e um relativo balanceamento entre as classes real e falso. Ele possui 72134 entradas, onde

35028 são reais e 37106 são falsas. No entanto, como é mostrado na seção 3.2.1, há duplicatas, o que torna o número de entradas válidas no *dataset* menor.

O *dataset* possui como colunas o identificador de cada texto, título, texto e *label*, onde a *label* 0 indica notícia falsa e 1 indica notícia real. Um ponto importante de ser notado é que, como o *dataset* não possui *features* pré-extraídas (apenas possui o texto em si), elas foram posteriormente extraídas pelos discentes (seção 3.2.1).

Uma análise inicial indica a seguinte proporção de duplicatas:

Duplicatas em títulos	9786
Duplicatas em textos	9415
Duplicatas em linhas inteiras	8456

Table 1. Quantidade de duplicatas encontradas na base de dados

Para este trabalho, considerou-se que apenas duplicatas tanto em título quanto em textos deveriam ser removidas. Alguns gráficos que mostram a distribuição dos tamanhos dos textos e títulos seguem.

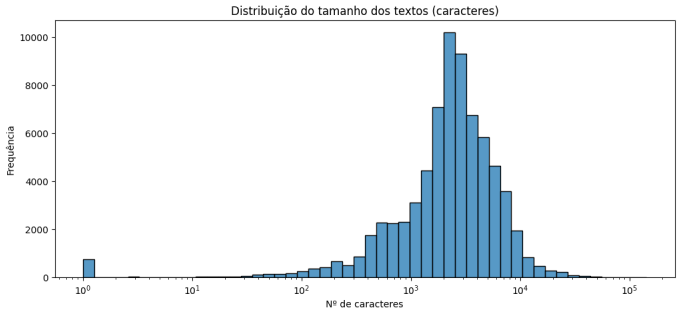

Figure 1. Distribuição dos tamanhos dos textos (em caracteres)

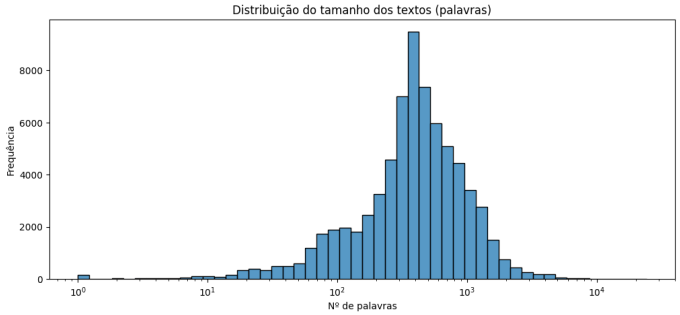

Figure 2. Distribuição dos tamanhos dos textos (em palavras)

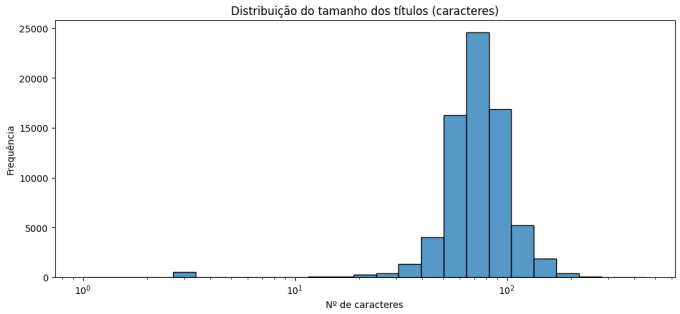

Figure 3. Distribuição dos tamanhos dos títulos (em caracteres)

3.1.2. Scikit-Learn

3.1.3. Torch

3.2. Métodos

3.2.1. Pré-processamento

3.2.2. Método 1: Técnicas clássicas

3.2.3. Método 2: Redes neurais

c. Metodologia contendo uma descrição das características da base de dados selecionada, os métodos aplicados, bem como da métrica de avaliação utilizada para verificar a qualidade dos métodos aplicados;

4. Resultados

5. Conclusões

References