Universidade Federal de São Carlos
Aprendizado de Máquina
Prof. Tiago A. Almeida <talmeida@ufscar.br>

Grupo 6: Detecção de *fake news* em notícias brasileiras

Problema

Com a popularização do acesso a *internet*, a facilidade de produção e consumo de conteúdo cresceu de forma muito rápida, resultando em profundas mudanças dentro de diversas áreas.

Uma das áreas mais afetadas dentro desta nova realidade foi o jornalismo. Através de *sites* e portais, as agências jornalísticas passaram a divulgar notícias de forma muito mais veloz, alcançando um número cada vez maior de leitores. Também foram vistas mudanças na linguagem utilizada e na relação do leitor com a notícia, devido principalmente ao advento das redes sociais e da difusão da jornalismo entre todas as esferas sociais.

Entretanto, devido a fatores como a possibilidade de lucro *online*, a criação de uma cultura de comunicação rápida dentro de redes sociais e, principalmente, a polarização política, deu-se início a um fenômeno de produção de *fake news*.

O termo, popularizado durante a eleição americana de 2016, caracteriza qualquer notícia falsa ou enganosa. Tais notícias são normalmente produzidas e divulgadas por *sites* de imprensa marrom e usuários de redes sociais, e já mostraram deter de um elevado poder de persuasão.

O problema está bastante presente na realidade do povo brasileiro, tendo crescido especialmente durante as eleições de 2018. E dentro de um mar de informações e notícias circulando pela *internet*, torna-se difícil para seus usuários saberem o que é verdadeiro e o que é falso.

Diante deste cenário, um núcleo de pesquisa montou uma base de notícias e rotulou-as como sendo verdadeiras ou *fake news*. O desafio proposto é a criação de um modelo capaz de processar notícias e classificá-las de acordo com a veracidade da informação passada.

Detalhes

Os conjuntos de dados, assim como demais informações sobre os mesmos, estão disponíveis em `https://github.com/roneysco/Fake.br-Corpus`.

Estão disponibilizadas duas pastas com os mesmos dados, processados de formas diferentes. O grupo deve trabalhar com o diretório *full_texts*, que contém os dados completos, sem nenhum processamento prévio, simulando assim o problema real.