UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
DEPARTAMENTO DE INFORMÁTICA
COLEGIADO DE CIÊNCIA DA COMPUTAÇÃO

ESTUDO DIRIGIDO EM INTELIGÊNCIA ARTIFICIAL: PARTE 1

MACHINE LEARNING

ATHUS ASSUNÇÃO CAVALINI

VITÓRIA, ES
2022

1. INTRODUÇÃO

2. OBJETIVO

O objetivo deste estudo dirigido em *Machine Learning* é analisar a aplicação de técnicas de aprendizado de máquina para analisar as mensagens que circulam no aplicativo Telegram relacionadas à Covid-19 e classificá-las como potenciais unidades de desinformação online.

Para o cumprimento do objetivo geral deste trabalho, é necessário cumprir os seguintes objetivos específicos:

a) Encontrar uma base de dados classificada (e preferencialmente validadas cientificamente) que possa relacionar mensagens online à desinformação, necessária para treinamento do algoritmo;
b) Realizar um pré-tratamento dos dados a fim de melhorar os resultados e desempenho dos algoritmos de classificação;
c) Comparar os resultados da aplicação de diferentes algoritmos para escolher aquele com resultados de maior qualidade;
d) Definir uma base de dados de mensagens de Telegram relacionadas à Covid-19;
e) Realizar o pré-tratamento das mensagens conforme o realizado na base de dados de treino;
f) Aplicar o algoritmo treinado para que classifique as mensagens e analisar os resultados obtidos;

3. METODOLOGIA

3.1 BASE DE DADOS E TRABALHOS RELACIONADOS

A base de dados utilizada para treino foi recuperada do repositório de dados de pesquisa do Departamento de Computação da Universidade Federal do Ceará (UFC) em parceria com a Universidade Estadual do Ceará (UEC).

Por conta da da recente popularização do Telegram, principalmente no Brasil, ainda são poucos os dados e produções científicas ao redor do mesmo. Dessa forma, a base de dados utilizada é composta de mensagens compartilhadas no aplicativo WhatsApp.

O *dataset* é composto por 2899 mensagens com uma média de 120 palavras cada. Ele foi classificado manualmente pela equipe de pesquisadores, quando 26% das mensagens foram consideradas contendo desinformação. Maiores informações podem ser consultadas no artigo[1].

Outro repositório, o FakeWhatsApp.BR[2], também realizou uma grande coleta e classificação de mensagens de WhatsApp (não necessariamente relacionadas à Covid-19).

Depois da classificação manual, aplicou-se classificações automáticas combinando diferentes técnicas de processamento e de aprendizado de máquinas. Com um total de 108 diferentes cenários, a melhor pontuação atingida pelos pesquisadores foi de 0.71.

No *dataset* escolhido, os pesquisadores também aplicaram técnicas de *Deep Learning* para classificar as mensagens e atingiram a pontuação máxima de de 0.774, justificada pelo tamanho reduzido das mensagens em comparação com os dados do repositório FakeWhatsApp.BR. Mais informações podem ser obtidas no artigo[3]. Filtradas as mensagens com menos de 50 palavras, a pontuação alcançada foi de 0.85.

3.2 TRATAMENTO INICIAL DOS DADOS

Os dados foram tratados da seguinte forma (e na seguinte ordem):

a) remoção de vírgulas, acentos e capitalização;
b) separação dos textos em palavras;
c) remoção de palavras com 3 ou menos caracteres;
d) remoção dos nomes de usuários, por se tratar de um treinamento que será aplicado aos dados de outra rede social;
e) remoção das *stopwords*;
f) tratamento das URLs, mantendo apenas o *hostname*;
g) e remoção dos números.

3.3 TREINO E AVALIAÇÃO DOS MÉTODOS

Foram testados 4 diferentes métodos de classificação: *OneVsOne, OneVsRest, Multinomial e AdaBoost*, todos disponibilizados na biblioteca *sklearn*. Para avaliar a

[1] Disponível em: https://sol.sbc.org.br/index.php/dsw/article/view/17422/17258
[2] Disponível em: https://github.com/cabrau/FakeWhatsApp.Br
[3] Disponível em: https://sol.sbc.org.br/index.php/sbbd/article/view/17868/17702

pontuação de cada um, foi utilizado o método *Cross Validation* (*cross_val_score*), também da biblioteca anteriormente citada.

Os métodos também foram aplicados nos dados a partir de variações do tratamento inicial (isto é, com e sem a retirada das stopwords ou dos números, por exemplo).

Por fim, o OneVsOne obteve melhores resultados, 81.55%. Por possuir uma execução relativamente rápida, manteve-se também o método Multinomial (77.75%) para comparação dos resultados. O método AdaBoost, por sua vez, possui tempo de execução até 10x maior e resultados próximos aos do primeiro (80.34%).

3.4 BASE DE DADOS REAL

Os dados a serem classificados foram retirados do dataset utilizado pelo autor em seu trabalho de conclusão de curso, que consiste em cerca de 3 milhões de mensagens enviadas no Telegram entre 01 de Agosto e 30 de Setembro de 2021.

O *dataset* compreende aproximadamente 300 canais e grupos, portanto foi realizado um agrupamento do conjunto utilizando-se a medida de *Modularity Class* considerando-se a taxa de encaminhamento de mensagens entre os grupos.

Dos 7 subconjuntos encontrados, um deles foi formado majoritariamente por canais e grupos cujo tema principal de discussão é a Covid-19 e temas afins, como vacina e medicina. Deste, foram selecionados 18 canais e grupos cujo título ou descrição fazia referência aos temas.

Do subconjunto, foram recuperadas as mensagens com mais de 30 caracteres, resultando em 89517 itens.

4. RESULTADOS

Por conta das limitações de tempo e memória, uma amostra aleatória de 10% das mensagens foi selecionada para ser classificada a partir dos algoritmos, isto é, 8951 mensagens.

Depois de executados os métodos, as mensagens classificadas foram exportadas para que se pudesse realizar uma validação manual. Mais uma vez por conta do tamanho, uma amostra aleatória foi separada, sendo 20 mensagens classificadas como *True* (isto é, possui desinformação) e 20 como *False*.

Os resultados foram:

Método	Falsos Positivos	Falsos Negativos	% de Acerto
OneVsOne	40%	20%	70%
Multinomial	10%	20%	85%
AdaBoost	40%	20%	70%

Dessa forma, pode-se concluir que apesar de não obter os melhores resultados na execução do Cross Validation, o modelo Multinomial aparentou apresentar uma menor taxa de falsos positivos (e consequentemente uma maior taxa de acerto) na classificação final.

Neste contexto, vale pontuar que existe uma diferença importante entre as taxas de falso positivo e negativo. Isso porque a maioria das redes sociais, como o Facebook, Instagram e Twitter, por exemplo, executa uma verificação manual dos conteúdos com potencial desinformação (seja por equipe própria ou em parceria com agências de checagem, as *fact-checking*).

Dessa forma, os falsos positivos são validados novamente, o que abre margem para uma taxa um pouco maior, enquanto os falsos negativos permitem que o conteúdo com desinformação circule por mais tempo nas redes e atinjam mais pessoas.

Todos os métodos apresentaram a mesma taxa de falsos negativos, de 20%. Isso significa que, a cada 10 mensagens não classificadas como desinformação pelo algoritmo, 2 deveriam ser. Isso mostra também a importância da possibilidade de se denunciar conteúdos nas plataformas, como aconteceu com o Twitter na última semana[4], de forma que essa taxa seja reduzida ainda mais.

[4] Disponível em: https://g1.globo.com/tecnologia/noticia/2022/01/17/twitter-adiciona-opcao-para-denunciar-fake-news-sobre-a-pandemia-no-brasil.ghtml