Projeto de Formatura – 2022

PCS - Departamento de Engenharia de Computação e Sistemas Digitais

Engenharia de Computação

Tema: **Racismo Amarelo: uma Análise sobre Discursos de Ódio Contemporâneos por meio de Aprendizagem de Máquina**

Contexto/Motivação

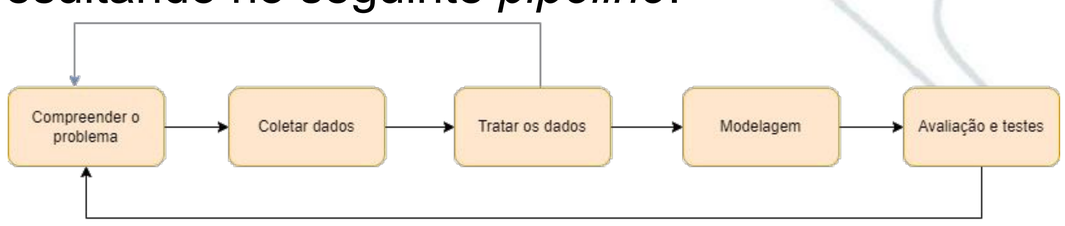
Na pandemia de Covid-19 e nas Olimpíadas de Tóquio, viu-se o aumento de casos de violência contra pessoas amarelas. A coalizão *Stop AAPI Hate* relatou 6603 ataques contra asiáticos entre março de 2020 e março de 2021, tanto verbais quanto físicos, em comércios, transporte público e locais de trabalho nos EUA. Pensando nisso, o projeto visa dar visibilidade ao tema e conscientizar o público geral sobre o racismo contra asiáticos e descendentes pelo Twitter.

Objetivo

O objetivo do projeto é, por meio da aplicação de estudos de machine learning e processamento de linguagem natural, classificar se mensagens do Twitter possuem conotação racista, com foco em povos asiáticos. Um bot na rede social é capaz de identificar tais mensagens ofensivas e auxiliar na sensibilização sobre o tema.

Metodologia

O desenvolvimento do projeto inspirou-se no TDSP (*Team Data Science Process*), metodologia iterativa para Data Science e Machine Learning, resultando no seguinte *pipeline*:



Foram treinados diferentes modelos de machine learning, com tweets coletados pela Twitter API, a partir das ofensas mais comuns usadas contra asiáticos, e do *dataset* público "Covid Hate".

Termos usados no filtro de coleta de tweets	
Termos	chink; gook; gookette; mongoloid; goloid; whoriental; rice nigger; dog eater; ching chong; slant eye; chinaman; zipperhead; jap/japs; butterhead; asian; churka; niakoue; korean; chinese; japanese
Hashtags	#ChinaDidThis, #ChinaLiedPeopleDied, #FuckChina, #MakeChinaPay, #WashTheHate, #BeCool2Asians, #StopAAPIHate, #ActToChange, #VeryAsian

Em paralelo ao estudo, desenvolvimento e treino dos modelos, houve a concepção do bot para o Twitter e a análise das limitações impostas pela plataforma a contas automatizadas.

Integrantes: Gabriel Oga Sanefuji
Sandra Ayumi Nihama

Professor Orientador: Ricardo Luis de Azevedo da Rocha

Resultados

Através dos diferentes testes realizados com as arquiteturas propostas, foi possível identificar os melhores modelos através do cálculo de métricas de performance:

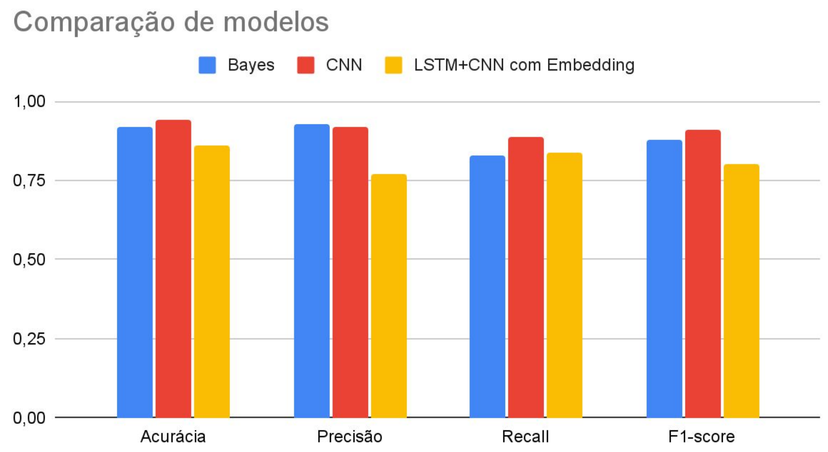


Figura 1: gráfico comparativo de três dos modelos treinados

O modelo de Naive Bayes, usado como *baseline* do projeto, foi superado apenas pelo modelo CNN, com acurácia de 0,94 e F1-score de 0,91.

Aplicação

Os modelos estão disponíveis para serem testados pelo público geral no site do projeto. Além disso, uma conta automatizada no Twitter é capaz de identificar comentários racistas em *threads* na plataforma, e pode ser ativada mencionando-a em comentários.

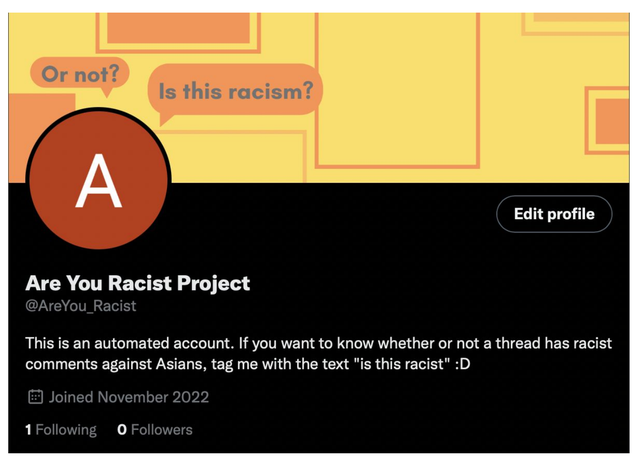


Figura 2: Imagem da conta automatizada do Twitter

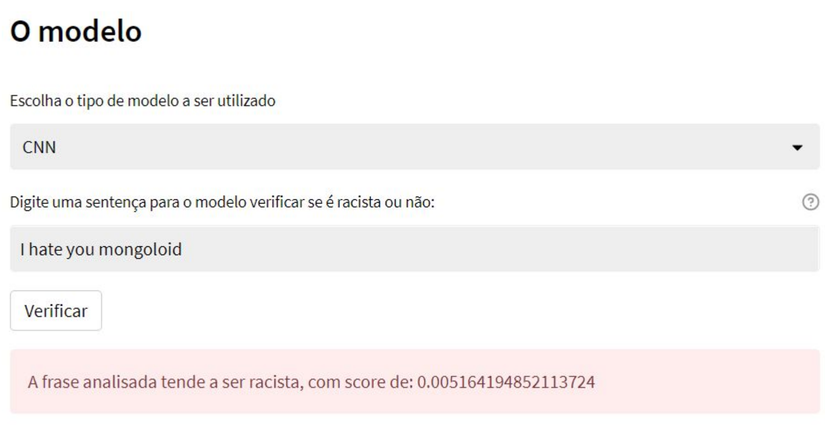


Figura 3: Imagem da tela de teste do modelo no site

[1] Microsoft. What is the Team Data Science Process?. Disponível em:https://learn.microsoft.com/en-us/azure/architecture/data-science-process/overview. Acesso em: 01 dez. 2022.

[2] B. He et al. “Racism is a virus: anti-asian hate and counterspeech in social media during the COVID-19 crisis”. Em: ASONAM ‘21: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 7 (nov. de 2021), pp. 90–94.