

DESENVOLVIMENTO DE UM APLICATIVO PARA CLASSIFICAR SENTIMENTOS DE TWEETS EM TEMPOS DE PANDEMIA

HELIKS MOTA MERSENBURG¹; LARISSA A. DE FREITAS²

¹Universidade Federal de Pelotas – hmmersenburg@inf.ufpel.edu.br

²Universidade Federal de Pelotas – larissa@inf.ufpel.edu.br

1. INTRODUÇÃO

A pandemia do coronavírus de 2019 (COVID-19) é considerada uma emergência de saúde pública de interesse internacional e representa um desafio para a resiliência psicológica. Dados de pesquisas são necessários para desenvolver estratégias baseadas em evidências para reduzir os impactos psicológicos adversos e os sintomas psiquiátricos durante a pandemia. Conforme apresentado por Wang *et al.* (2020) em uma pesquisa realizada na crise da COVID-19, verificou-se que, dentre 1210 participantes, 53,4% apresentaram sequelas psicológicas moderadas ou severas, incluindo sintomas depressivos (16,5%), ansiedade (28,8%) e estresse (8,1%).

Conforme citado em Calvo *et al.* (2017), as técnicas de Processamento de Língua Natural (PLN) podem ser usadas para fazer inferências sobre os estados mentais das pessoas a partir do que elas escrevem nas redes sociais. Essas inferências podem então ser usadas para criar caminhos online para direcionar as pessoas para informações e assistência de saúde e também para gerar intervenções personalizadas.

A área de Análise de Sentimento é o estudo computacional de opiniões, sentimentos e emoções que estão expressas em um texto, a classificação de texto pode ser baseada em aprendizado de máquina, regras ou em formato híbrido conforme apresentado por Liu (2012). Uma vez que é um problema de classificação de texto, qualquer método de aprendizagem supervisionada existente pode ser aplicado (LIU, 2012). O Aprendizado de Máquina supervisionado necessita de dados previamente rotulados para formar uma base de treinamento e construir uma função que mapeia dados de entrada para saída com base nos pares de dados de treinamento (RUSSELL, 2019).

Neste trabalho propomos desenvolver um modelo de Análise de Sentimentos (AS) baseado em Aprendizado de Máquina (AM) com o intuito de criar um aplicativo para dispositivos móveis que classifica sentimentos de tweets escritos durante a COVID-19. O aplicativo irá classificar os *tweets* de determinado usuário e apresentar a porcentagem de *tweets* com tendências ansiosas, depressivas ou sem tendência aparente. Além disso, o aplicativo também descreverá e apresentará campanhas do governo ou de instituições não governamentais que estejam ligadas ao tratamento e prevenção de causas psicológicas, assim como as hashtags mais utilizadas em campanhas no Twitter¹.

2. METODOLOGIA

Em um primeiro momento, realizamos uma revisão bibliográfica, ou seja, buscamos no Google Scholar² por trabalhos sobre AS baseado em AM, tanto em

¹ <https://twitter.com/>

² <https://scholar.google.com.br/>

português quanto em inglês, aplicados a textos sobre transtornos depressivos, preferencialmente em tempos de pandemia.

Os trabalhos relacionados foram divididos em duas partes. A primeira refere-se a trabalhos sob revisão de escopo, ou seja, trabalhos que mapeiam os conceitos-chave, os quais requerem uma compreensão da literatura existente e suas lacunas envolvendo um processo que investiga uma questão de pesquisa que desenvolve e aplica uma estrutura para explorar questões utilizando a literatura existente (ARMSTRONG *et al.*, 2011). A segunda refere-se a trabalhos que foram publicados em tempo de pandemia, selecionamos apenas trabalhos que abordam AM sobre o assunto Saúde Mental (depressão, ansiedade, entre outros). Os algoritmos *Support Vector Machine* (SVM) e *Naïve Bayes* (NB) foram os mais utilizados nos trabalhos relacionados, por este motivo e pelo fato do algoritmo SVM apresentar melhor Medida-F no *dataset* elaborado por (LOW *et al.*, 2020), optamos por utilizar as duas técnicas de AM (SVM e NB) no presente trabalho.

Para criar um modelo de AS baseado em AM utilizaremos o *dataset* de Low *et al.* (2020). Vale ressaltar que este *dataset* foi coletado inteiramente da rede social Reddit³ e contém um conjunto de dados formado por 28 subreddits dos quais 15 deles estão associados ao apoio à saúde mental, coletados no ano de 2018 a 2020, possuindo 327289 postagens feitas antes da pandemia e durante a pandemia.

Na Figura 1 temos o fluxograma contendo os passos que serão percorridos pelo modelo. Iniciamos o fluxo de desenvolvimento aplicando as técnicas de pré-processamento no *dataset* elaborado por (LOW *et al.*, 2020). Pelo fato de utilizarmos a técnica supervisionada para obtenção de resultados, dividimos os dados em duas partes, a primeira parte, contendo 75% dos dados de treinamento, que contém o par de objetos (vetor de entrada e um valor de saída), e o restante, contendo 25% dos dados de teste, que contém apenas o vetor de entrada. Na mesma figura apresentamos a coleta de tweets utilizando a API Tweepy⁴. Estes dados correspondem aos dados de teste, os quais serão pré-processados e após uso do modelo de classificação será feita a previsão da emoção/sentimento (por exemplo: ansiedade e depressão) presente no mesmo.

3. RESULTADOS E DISCUSSÃO

Até o momento já implementamos algumas técnicas de pré-processamento tais como: remoção de *links*, remoção de caracteres especiais, uso de tradução automática, remoção de *stopwords*, filtragem de postagens com 240 caracteres e detecção de entidades nomeadas, também implementamos o coletor de *tweets* por intervalo de datas e usuário.

As figuras 2 e 3 apresentam um esboço das telas *home*, notícias, perfil e campanhas da aplicação. Algumas delas já foram implementadas (*home* e notícias) e outras estão em andamento (perfil e campanhas).

4. CONCLUSÕES

Em suma, nosso objetivo é implementar uma interface simples e funcional para que com poucos cliques o usuário consiga visualizar a classificação de

³ <https://www.reddit.com/>

⁴ <https://www.tweepy.org/>

emoção/sentimento (depressão e ansiedade) de seus próprios tweets em um período de tempo. O segundo objetivo é criar um modelo de AM usando diferentes técnicas de pré-processamento para encontrar resultados promissores para utilizar na nossa aplicação. Com estes dois pontos bem fundamentados, tentaremos garantir que o usuário tenha uma boa experiência de uso da ferramenta e que busque apoio caso seja necessário.

Figura 1 – Fluxo do treinamento, teste e background da aplicação. Fonte: Própria.

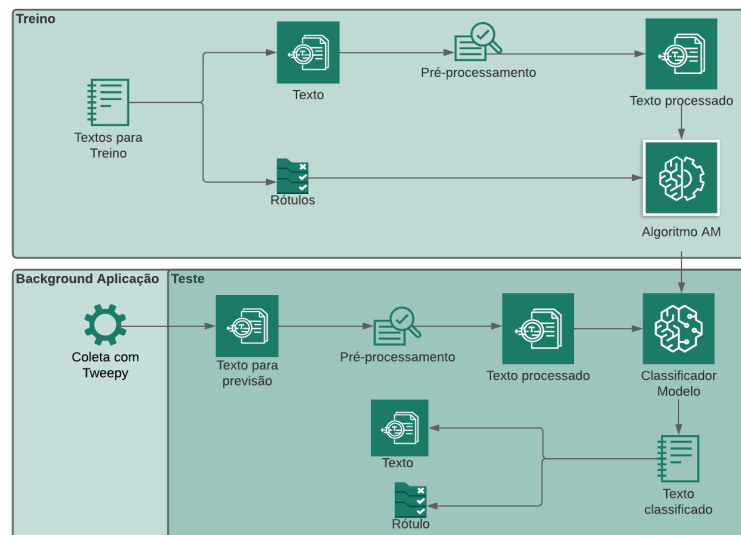
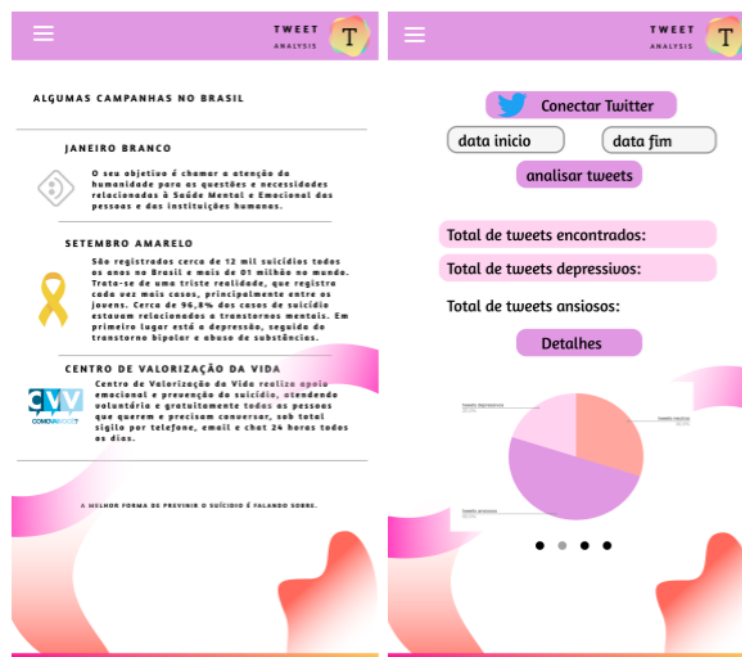


Figura 2 – Esboço das telas *home* e *notícias* da aplicação. Fonte: Própria.



Figura 3 – Esboço das telas *Campanhas* e *Perfil* da aplicação. Fonte: Própria.



5. REFERÊNCIAS BIBLIOGRÁFICAS

WANG, G.W.; ZHANG, Y.Z.; ZHAO, J.Z.; ZHANG, J.Z.; JIANG, F.J. Mitigate the effects of home confinement on children during the COVID-19 outbreak. **The Lancet**, Amsterdam, v.395, n.10228, p.945-947, 2020.

CALVO, R.A.C; MILNE, D.N.M.; HUSSAIN, M.S.H.; CHRISTENSEN, H.C. Natural language processing in mental health applications using non-clinical texts. **Cambridge University Press**, New York, v.23, n.5, p.649-685, 2017.

ARMSTRONG, R.A.; HALL, B.J.H.; DOYLE, J.D.; WATERS, E.W. 'Scoping the scope' of a cochrane review. **Journal of Public Health**, New York, v.33, n.1, p.147-150, 2011.

LOW, D.L.; RUMKER, L.R.; TALKAR, T.T.; TOROUS, J.T.; CECCHI, G.C.; GHOSH, S.G. Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. In: **JOURNAL OF MEDICAL INTERNET RESEARCH**, 22., Toronto, 2011. **Anais...** Toronto: JMIR Publications, 2011. v.22. p.22635.

LIU, B.L. Sentiment Analysis and Opinion Mining. **Synthesis Lectures on Human Language Technologies**, Williston, v.5, n.1, p.1-167, 2012.

RUSSELL, S.R. **Human compatible: Artificial intelligence and the problem of control**. Westminster: Penguin, 2019.