RESUMO - TÉCNOLOGIA EM DESENVOLVIMENTO DE SISTEMAS E TECNOLOGIA DA INFORMAÇÃO

PIPELINE DE PRÉ-PROCESSAMENTO E EXTRAÇÃO DE CARACTERÍSTICAS PARA CLASSIFICAÇÃO DE SENTIMENTOS EM TEXTOS DE SAÚDE MENTAL EM REDES SOCIAIS

Bruna Pisani De Souza (anyapisani@gmail.com)

Dayane Perez Bravo (dayane.b@uninter.com)

Ederson Cichaczewski (ederson.c@uninter.com)

Fabricio Jorge Souza Magalhães (fabricioj.dev@gmail.com)

O Processamento de Linguagem Natural (NLP) é um ramo da Inteligência Artificial e da Linguística, dedicado a fazer com que os computadores compreendam declarações ou palavras escritas em linguagem humana. Dentro do âmbito da saúde mental o NLP tem um papel de destaque, sendo uma abordagem promissora para identificar padrões emocionais que podem contribuir para a detecção precoce de sinais de sofrimento psíquico em seres humanos. Este trabalho propõe uma metodologia para analisar sentimentos em posts do subreddit “Conversas” na rede social Reddit, usando técnicas computacionais de análise de sentimentos, através de abordagens de inteligência artificial. A coleta de dados foi feita por meio de uma API pública do Reddit, respeitando a LGPD com anonimização dos usuários e as diretrizes éticas para uso de dados em pesquisas. A linguagem de programação empregada foi Python, juntamente com bibliotecas como PRAW, Pandas, Langdetect, Emoji e RE, utilizadas para o pré-processamento e a organização

eficiente dos dados. A coleta foi dividida em três categorias de sentimento: positivo, negativo e neutro, classificadas com base em critérios pré-estabelecidos. Para cada categoria, foram selecionadas palavras-chave representativas, como "amo" e "feliz" para o sentimento positivo, ou "triste" e "ódio" para o negativo. Essas palavras foram analisadas com a biblioteca WordCloud, responsável pela geração de nuvens de palavras. Essa técnica permite uma familiarização rápida com o conteúdo de grandes coleções textuais, identificando domínios temáticos em poucos segundos. Etapas de pré-processamento foram necessárias para a geração de uma nuvem de palavras, considerando a presença marcante de gírias, expressões informais e regionalismos no corpus. Inicialmente, a filtragem de stop words em português foi realizada com o auxílio da biblioteca NLTK, porém sem melhorias significativas na qualidade do resultado. Em uma segunda tentativa, adotou-se uma abordagem que combinava uma lista manual de termos frequentes com a métrica TF-IDF (Term Frequency - Inverse Document Frequency), a qual pondera as palavras conforme sua frequência relativa no conjunto de documentos. Apesar disso, os resultados permaneciam insatisfatórios. Diante das limitações encontradas, e com base na metodologia de análise textual, implementou-se um pipeline de pré-processamento mais robusto, composto pelas etapas de tokenização, etiquetagem morfossintática (POS tagging), executadas com a biblioteca SpaCy – e lematização, esta última utilizando o modelo linguístico mais abrangente (550MB) para o português. No processo, foi mantida a existência da lista manual de exclusão e do filtro do NLTK. Essa abordagem multidimensional resultou finalmente, em nuvens de palavras com excelente definição. Para as próximas etapas da pesquisa, será feita a extração de features dos textos processados, que serão subsequentemente submetidos à classificação por meio de algoritmos de aprendizado de máquina clássicos, como Random Forest, Naive Bayes, SVM (Máquinas de Vetores de Suporte) e Regressão Logística. Os resultados obtidos apresentaram potencial de contribuir para o desenvolvimento de ferramentas de suporte à saúde mental, auxiliando profissionais e usuários na detecção precoce de sinais de sofrimento psíquico e na promoção do bem-estar emocional.

Palavras-chave: processamento de linguagem natural; análise de sentimentos; saúde mental; reddit;.