Ricardo da Silva Santos
IMECC/UNICAMP
r263150@dac.unicamp.br

Cristiano Torezzan
FCA/UNICAMP
torezzan@unicamp.br

RESUMO

O Processamento de Linguagem Natural (NLP, do inglês Natural Language Processing) é uma área da inteligência artificial que se dedica ao desenvolvimento de métodos que possibilitem aos computadores entender, interpretar e gerar linguagem humana natural. É uma área interdisciplinar que combina conhecimentos de computação, linguística, psicologia e áreas específicas do domínio de aplicação. NLP é usado em uma ampla variedade de aplicações como sistemas de chatbot, análise de sentimentos em redes sociais, tradução automática, reconhecimento de fala e escrita, geração automática de texto, entre outras. Na área médica, uma importante aplicação de NLP consiste na extração de dados a partir de textos não estruturados escritos em linguagem natural. Para isso, duas técnicas tem sido particularmente utilizadas: Expressões Regulares (RegEx), que consiste no uso de expressões padronizadas para a criação de padrões de pesquisa, que são utilizados para buscar (minerar) determinados campos específicos em textos [1]. Reconhecimento de Entidades Nomeadas (NER, do inglês Named Entity Recognition), também chamado de fragmentação, extração ou identificação de entidades, é uma sub-área de NLP que se dedica à tarefa de identificar e categorizar as principais informações (entidades) em um texto. Uma das principais aplicações de NER é a extração de informações específicas a partir de textos não estruturados. Na área da saúde, essa abordagem pode ser utilizada para predizer diagnósticos, ou códigos de CID (Classificação Internacional de Doenças). Recentes avanços da literatura tem mostrado que modelos pré-treinados podem alcançar resultados com ótimo desempenho para textos em língua inglesa [2] [3] e já existe adapatações desse modelo para lingua coreana[4]. Neste trabalho apresentamos alguns resultados preliminares de uma proposta que utiliza os modelos BIO-BERT e Clinical-BERT para associação de sintomas a CID em língua portuguesa utilizando a biblioteca pré-treinada BERTimbau [5].

PALAVRAS CHAVE. BERT, PLN, CID.

Tópicos (SA – PO na Área de Saúde)

Referências

[1] M. Cui, R. Bai, Z Lu, X. Li, U. Aickelin, and P. Ge. *Regular Expression Based Medical Text Classification Using Constructive Heuristic Approach*. IEEE Acess 7, 2019.

[2] E. Alsentzer, J. Murphy, W. Boag, W.H. Weng, D. Jin, T. Naumann, and M. McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 2019.

[3] E. Al-Bashabsheh, A. Alaiad, M. Al-Ayyoub, et al. Improving clinical documentation: automatic inference of icd-10 codes from patient notes using bert model. *Journal of Supercomputing*, 2023. doi: 10.1007/s11227-023-05160-z.

[4] Sangah Lee, Hansol Jang, Yunmee Baik, Suzi Park, and Hyopil Shin. Kr-bert: A small-scale korean-specific language model. *ArXiv*, abs/2008.03979, 2020.

[5] F. Souza and R. Nogueira, Lotufo. *BERTimbau: Pretrained BERT Models for Brazilian Portuguese*. Intelligent Systems. BRACIS 2020. Lecture Notes in Computer Science, vol 12319. Springer, 2020.