World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:18, No:05, 2024

Benchmarking Bert-Based Low-Resource Language: Case Uzbek NLP Models

Authors : Jamshid Qodirov, Sirojiddin Komolov, Ravilov Mirahmad, Olimjon Mirzayev

Abstract : Nowadays, natural language processing tools play a crucial role in our daily lives, including various techniques with text processing. There are very advanced models in modern languages, such as English, Russian etc. But, in some languages, such as Uzbek, the NLP models have been developed recently. Thus, there are only a few NLP models in Uzbek language. Moreover, there is no such work that could show which Uzbek NLP model behaves in different situations and when to use them. This work tries to close this gap and compares the Uzbek NLP models existing as of the time this article was written. The authors try to compare the NLP models in two different scenarios: sentiment analysis and sentence similarity, which are the implementations of the two most common problems in the industry: classification and similarity. Another outcome from this work is two datasets for classification and sentence similarity in Uzbek language that we generated ourselves and can be useful in both industry and academia as well.

Keywords : NLP, benchmak, bert, vectorization
Conference Title : ICCLNLP 2024 : International Conference on Computational Linguistics and Natural Language Processing
Conference Location : Singapore, Singapore
Conference Dates : May 02-03, 2024