Generating synthetic text samples to address class imbalance in NLP

We are looking for a master/bachelor student who is interested in text classification and text generation to work under supervision of Yousef Younes. You will be working with us on the task of classifying whether a scientific text contains dataset mentions or not as part of the DFG project UnknownData. Dataset mentions can refer to the actual data used in the paper or just point to some dataset in the literature. These mentions are important because datasets are essential to reproduce results and to compare model's performance. An important challenge for us is that we don't have enough annotated examples, so our idea is to automatically generate more of them using text generation [1] for training AI. Since the literature is biased towards the texts that do not contain dataset mentions, the focus will be to alleviate this bias by generating texts that contain dataset mentions [2]. Then the classification performance will serve as an indirect evaluation for the quality of the generated samples.

For further information concerning the task please contact Yousef Younes via yousef.younes@gesis.org.

[1] Iqbal, Touseef, and Shaima Qureshi. "The survey: Text generation models in deep learning." Journal of King Saud University-Computer and Information Sciences (2020).
[2] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.