Online Bert-based Topic Modelling

Eric Gericke and **Nicolas Jamet** and **Tian Guo** and **Martin Schüle**
geci@zhaw.ch

Abstract

Topic modelling is a machine learning method for iden- tifying and extracting relevant topics from large amounts of text. It allows patterns and trends to be identified in the data that would otherwise be difficult to spot and large amounts of unstructured data to be organized. With news data, for example, it enables relevant topics to be identified and anal- ysed in real time, helping to find information on specific topics quickly and easily. There are a number of different types of topic models such as the classic Latent Dirichlet Allocation (LDA) [1] and variants thereof. In recent years BERT-based models have become popular as they provide a more accurate and contextual text representation allowing for improved topic identification and categorisation. With so-called dynamic topic modelling we can track topics over time thereby analysing not only the current top- ics but also their development over time. Thus, dynamic topic modelling may enable a better understanding of com- plex temporal patterns and can be used to predict future trends and topic developments. However, the constant flux of new data, e.g. in news stream settings, poses a challenge as the standard dynamic topic modelling methods usually need to use the entire dataset resulting in significant difficulties as data volumes grow. The solution to this problem is to move to online models that are able to incrementally process and integrate new data, thereby avoiding the need to retrain the model with the entire dataset. In this contribution we demonstrate a BERT-type dy- namic topic modelling approach which can reliably track topics over time without the need to merge datasets or to re-train models on the full dataset.