Classification algorithms in text analytics

Tutor	Luigi Curini is Professor of Political Science at University of Milan (Italy) and Visiting Professor at Waseda University (Tokyo). His research interests include party competition, legislative behavior, and text analytics. He is the co-editor of the SAGE Handbook of Research Methods in Political Science & International Relation (2020). His latest book is: Discussing the Islamic State on Twitter, London: Palgrave/MacMillan, 2022
Organization	Digital Skills, University of Lucerne
Language	English
ECTS-Points	4
Contact	samuel.schmid@unilu.ch
Dates and time	Friday 28 March 9:15-16:45 Saturday 29 March 9:15-16:45 Friday 4 April 9:15-16:45 Saturday 5 April 9:15-16:45 The course will be offered on-line.
Content	In this workshop we will introduce some of the classification techniques developed in the literature to deal with digital texts. The aim is to provide an introductory guide to this exciting area of research, while also offering guidelines on how to effectively use statistical methods on texts for social scientific research by discussing the advantages, but also the limits, of each approach. The attention will be devoted to three main areas:

	1) unsupervised classification that allow to discover new ways of organizing texts into a set of unknown categories; 2) supervised classification methods that allow to organize texts into a set of pre-defined categories (both via automatic as well as human-tagging, i.e., dictionary and Machine Learning algorithms); 3) word-embedding techniques that extend beyond the traditional bag-of-words approach in text analytics. We will explore how these techniques, along with their recent advancements based on the self-attention mechanism, underpin the revolution of Large Language Models (LLMs). Day one, Content: introduction to text analytics; unsupervised classification methods Day two, Content: supervised classification methods (dictionary and machine learning algorithms) Day three, Content: neural network models; validation and interpretation of machine learning algorithms Day four, Content: word-embedding techniques; an introduction to Large Language Models architecture
Prerequisites/Materials	An elementary knowledge of R (having attended any of the introductory workshops offered by the Digital Skills program usually satisfies this requirement), plus a curiosity towards applied statistics, are good prerequisites for the lab sessions. Participants will familiarize with several R packages, including quanteda, topicmodels, stm, naivebayes, randomForest, keras, text. All the datasets, replication files of the lab sessions and reference texts will be made available at a dedicated URL before the beginning of the workshop. Workshop participants should bring their own laptop with R, RStudio and the relevant packages previously installed and functioning (instructions will be circulated beforehand). Participants will also become familiar with using

	Google Colaboratory (Colab): a programming environment which allows the user to run code in the browser without the need to have installed R or RStudio in a laptop.