

## Text Classification Algorithms: From Bag-of-Words to Dynamic Word Embeddings

<b>Tutor</b>	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
<b>Organization</b>	Digital Skills Programme, University of Lucerne
<b>Language</b>	English
<b>ECTS-Points</b>	2
<b>Contact</b>	<a href="mailto:lumacss@unilu.ch">lumacss@unilu.ch</a>
<b>Dates and time</b>	In-person in Lucerne: Friday, 20 March 9:15-16:45, HS15 Saturday, 21 March 9:15-16:45, HS15 Online: Friday, 27 March 9:15-16:45 Saturday, 28 March 9:15-16:45
<b>Content</b>	In recent years, interest in text analysis within the social sciences has grown dramatically. This rise is largely driven by new techniques that enable researchers to draw substantively important inferences about politics and society from large text collections. This course provides students with the foundational skills needed to work with text data in their research, along with guidance on how to appropriately use and validate text-based methods in social scientific inquiry. It also serves as a gateway to more advanced study in artificial intelligence and large language models. We begin with the bag-of-words approach and then move on to static and dynamic word embeddings.

	<p>The course focuses on three main areas:</p> <ol style="list-style-type: none"> <li>1. <b>Supervised classification methods</b> (i.e., machine learning algorithms) used to categorize texts into predefined classes. We will introduce key concepts common to ML techniques—such as overfitting, cross-validation, and model interpretability—and discuss how to apply popular algorithms, including random forests, naïve Bayes, and neural networks.</li> <li>2. <b>Word-embedding techniques</b> that go beyond the traditional bag-of-words framework. We will examine both static and dynamic embeddings, with particular attention to methods that rely on self-attention mechanisms.</li> <li>3. <b>The role of dynamic word-embedding methods in the rise of Large Language Models (LLMs).</b> We will place special emphasis on the BERT family of Transformers, including how to use and fine-tune them. Approaches to Natural Language Inference (NLI) will also be discussed.</li> </ol> <p><b>Contents across course days:</b></p> <ul style="list-style-type: none"> <li>• Day one: introduction to text analytics and supervised classification methods</li> <li>• Day two: neural network models and validation of machine learning algorithms</li> <li>• Day three: interpretation of machine learning algorithms and word-embedding techniques</li> <li>• Day four: Large Language Models architecture</li> </ul>
<p><b>Prerequisites/Materials</b></p>	<p>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 <code>quantedanaivebayes</code>, <code>randomForest</code>, <code>keras3</code>, <code>lime</code>, <code>ingredients</code>, <code>DALEX</code>.</p> <p>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.</p>

	<p>Workshop participants should bring their own laptop with R, RStudio and the relevant packages previously installed and functioning (instructions will be circulated beforehand).</p> <p>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.</p>
--	---