



Course Code & Title	LISS2117 Quantitative methods for text classification and topic detection		
Convenor(s)	Dr Michele Scotto di Vettimo , Leverhulme Early Career Fellow, Department of Political Economy, King's College London		
Institution	King's College London	Department	Political Economy
Academic Year	2025-26	Term	Summer
Number of Sessions	3	Length of Session(s)	6 hours
	Day, Date	Start and End time	Room Location
	30/04/2026 05/05/2026 07/05/2026	10:00 – 16:00	Online
Enrolment Link:	Available to book on SkillsForge from 2 March 2026 . Click the course link to log in and register. Questions? Visit our Training FAQ here: Frequently Asked Questions - LISS DTP (liss-dtp.ac.uk)		

Meet your course convenor

Michele is a Research Associate/Leverhulme Early Career Fellow in the Department of Political Economy at King's College London. Previously, he worked as Research Fellow at the University of Exeter.

Michele obtained his PhD in Political Science from the King's College London and holds a BA in Political Science and MAs in Management of Public Administrations and European Affairs.

Michele is particularly interested in public support for European integration and the responsiveness of non-majoritarian institutions and of the EU policy-making process. His current project is about decision-making in the European Council.

Course Description

This course introduces PhD students to various methods for automated text classification for the social sciences. The methods covered allow, among other things, to automatically annotate texts (e.g., identify sentences with hate language in a large corpus), or to detect which topic they are about (e.g., to what extent a newspaper article talks about the economy).

The sessions cover approaches with different levels of sophistication, whilst also focusing on foundational and cross-cutting concepts that are relevant for all analyses relying on automated text classification.



The course is structured in three sessions. Beyond introducing various methods for automated text classification, each session also covers practical exercises to allow students to familiarise with the methods covered in the session.

Learning Outcomes

At the end of this course, students will be able to:

- Understand fundamental aspects of quantitative approaches to text classification, such as accuracy, validation, and reliability.
- Understand the functioning of various methods to perform automated text classification, their differences and their pros and cons.
- Apply these methods to their own text corpus to address a substantive research question.
- Critically evaluate social science research that uses automated text analysis methods for text classification.

Course Outline

1. Key concepts for automated classification and introduction to bag-of-words approaches

Session 1 covers foundational concepts related to automated text classification and introduces some basic quantitative approaches. Firstly, it puts text classification in the wider context of quantitative text analysis and clarifies its scope and relationship with other research focuses (e.g., text scaling). Secondly, it covers basic methods for automated classification such as word counts and dictionary methods. Thirdly, it focuses on general theoretical and practical aspects related to the validation of the results, which will be further elaborated throughout the course and tailored to the relevant text analysis method. Finally, it introduces foundational elements of bag-of-word approaches (e.g., tokens, document-feature matrices).

Required readings for Session 1:

- Grimmer, J., & Stewart, B. (2013). *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts*, *Political Analysis*, 21(3), 267-297.
- Grimmer, J., Roberts, M., & Stewart, B. (2022). *Text As Data. A New Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press. Chapters 5 and 16.

Other optional readings:

- Grimmer, J., Roberts, M., & Stewart, B. (2022). *Text As Data. A New Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press. Chapter 15.
- Van Atteveldt, W., Van der Velden, M. A., & Boukes, M. (2021). *The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms*. *Communication Methods and Measures*, 15(2), 121-140.

2. Topics models and machine-learning algorithms for classification

Session 2 further expands the discussion of bag-of-words approaches by covering topic models and machine-learning algorithms for text classification. Firstly, the session focuses on topic classification using semi-supervised topic models, which will also be compared to unsupervised alternatives, as well as to simpler



approaches covered in Session 1. Secondly, it introduces the logic of automated classification via machine learning and presents some widely used algorithms for text classification. In so doing, it expands on issues related to model training and validation of the results.

Required readings for Session 2:

- Grimmer, J., Roberts, M., & Stewart, B. (2022). *Text As Data. A New Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press. Chapters 13, 19, and 20.

Other optional readings:

- Anastasopoulos, L. J., & Bertelli, A. M. (2020). Understanding delegation through machine learning: A method and application to the European Union. *American Political Science Review*, 114(1), 291-301.
- Barberá, P., Boydston, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, 29(1), 19-42.
- Watanabe, K., & Zhou, Y. (2022). Theory-driven analysis of large corpora: Semisupervised topic classification of the UN speeches. *Social Science Computer Review*, 40(2), 346-366.

3. Word-embeddings approaches and large language models

Session 3 moves away from bag-of-words approaches and introduces novel methodologies based on word-embeddings and large language models. Firstly, it presents embeddings representation of words and their key properties. It then replicates some of the machine-learning algorithms covered in the previous session to show how they can handle both bag-of-words and word-embeddings representations. Secondly, this session focuses on large language models, particularly transformers models. It covers the use and fine-tuning of pre-trained models for text classification. Thirdly, it introduces natural language inference as a strategy for text classification. Finally, the session gives an overview of models capable of classifying texts in multilingual contexts.

Required readings for Session 3:

- Grimmer, J., Roberts, M., and Stewart, B. (2022). *Text As Data. A New Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press. Chapter 8.
- Laurer, M., van Atteveldt, W., Casas, A., & Welbers, K. (2023). Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis*, 32(1), 84-100.
- Rodriguez, P. L., & Spirling, A. (2022). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics*, 84(1), 101-115.

Other optional readings:

- Laurer, M., van Atteveldt, W., Casas, A., and Welbers, K. (2024). On Measurement Validity and Language Models: Increasing Validity and Decreasing Bias with Instructions, *Communication Methods and Measures*, 1-17.



- Rodman, E. (2020). A timely intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis*, 28(1), 87-111.

Reading list

N.B. This is a general background reading list (please see details under "Course Outline" on specific preparation for each session). The selection of material below is aimed at offering an overview of additional sources useful in terms of concepts, methodologies, and substantive applications.

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. G. (2022). Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, 16(1), 1-18.

Benoit, K. (2020). *Text as data: An overview*. In "The SAGE Handbook of Research Methods in Political Science and International Relations". London: SAGE Publishing.

Bernhard, J., Teuffenbach, M., & Boomgaarden, H. G. (2023). Topic Model validation methods and their impact on Model selection and evaluation. *Computational Communication Research*, 5(1), 1-26.

Birkenmaier, L., Lechner, C., & Wagner, C. (2024). The search for solid ground in text as data: A systematic review of validation approaches. *Communication Methods and Measures*, 8(3), 249-277.

Chen, Y., Peng, Z., Kim, S. H., & Choi, C. W. (2023). What we can do and cannot do with topic modeling: A systematic review. *Communication Methods and Measures*, 17(2), 111-130.

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political analysis*, 26(2), 168-189.

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society*, 2(2), 1-5.

Eshima, S., Imai, K., & Sasaki, T. (2024). Keyword-assisted topic models. *American Journal of Political Science*, 68(2), 730-750.

Licht, H. (2023). Cross-lingual classification of political texts using multilingual sentence embeddings. *Political Analysis*, 31(3), 366-379.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Haussler, T., Schmid-Petri, H., & Adam, S. (2021). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3), 13-38.

Reveillhac, M., & Morselli, D. (2022). Dictionary-based and machine learning classification approaches: a comparison for tonality and frame detection on Twitter data. *Political Research Exchange*, 4(1), 2029217.

Song, H., Tolochko, P., Eberl, J. M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgaarden, H. G. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication*, 37(4), 550-572.

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political communication*, 29(2), 205-231.

Widmann, T., & Wich, M. (2023). Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in German political text. *Political Analysis*, 31(4), 626-641.



Eligibility

There are no formal prerequisites to registering on this module. However, a very basic knowledge of the logic of programming languages for data analysis -- like R or Python -- is required (e.g., how to import data). Some basic statistical knowledge or would be advantageous. Automated text classification is employed in various disciplines, and this course could be interesting and useful for research students in many fields in the social sciences.

Number of students

Min: 1

Max: 20