Big Data and Social Media Research 1 (RECSM Summer School 2019)

Cornelius Puschmann, Leibniz Institute for Media Research
c.puschmann@hans-bredow-institut.de

Course description

Across global publics, social media platforms play an increasingly important role in politics, business, culture and academia. From services such as Facebook, Twitter, YouTube and Instagram in Western and Middle Eastern countries to platforms like VK (Russia) and WeChat (China), social media are used for a diverse set of purposes by a wide range of actors, from government entities and political activists to celebrities and public intellectuals. They also play a controversial role for public debate, having both been framed as instruments of democratization and openness and as dangerous, polarizing and pervaded by misinformation and extremism. What is largely undisputed however is that social media represent a vital data source for the study of politics, culture and society at large, and are therefore of growing relevance to empirical social science.

This class focuses on how the types of questions that are relevant to political science, sociology and communication studies may be approached using digital data from social media platforms in combination with innovative computational methods for content analysis ("big data" research). The platforms used as examples include Twitter, Facebook and YouTube, and the techniques covered will include sentiment analysis and supervised machine learning.The course will follow a hands-on approach, with short theoretical sessions followed by coding challenges where participants will need to apply new methods.

The first segment of the class will cover research design issues and the choice of appropriate methods and data, as well as practical considerations related to data acquisition, preprocessing and classification. Participants will be introduced to the quanteda package for R, which represents a useful framework for the analysis of textual data. They will also learn the basics of retrieving data through the Twitter API using the rtweet package.

The second segment will introduce participants to sentiment analysis and more broadly to so-called dictionary methods for studying issues in social media discourse. A number of broadly available resources such as the Lexicoder Policy Agendas and Laver-Garry dictionaries will be employed, along with the extensive LIWC lexicon.

The third segment will address supervised machine learning and demonstrate how this technique may be used to classify social media postings by a number of criteria, including sentiment. Participants will learn how to evaluate an existing manually performed classification and how to statistically judge the performance of different SML algorithms).

Aims

Participants will learn how to obtain and analyze large-scale social media data sets to answer questions relevant to their field. In order to achieve this goal, they will be introduced to the use of R for content analysis with quanteda and RTextTools. They will also learn the fundamentals of interacting with social media platform APIs, as well as managing data and visualizing results.

Prerequisites

The course will assume general familiarity with R. Ideally, participants should be able to know how to read datasets in R, work with vectors and data frames, and run basic statistical analyses, such as linear regression.

Software and data

The course will use the open-source software R and the development environment RStudio, which greatly facilitates coding with R. Both R and RStudio are freely available and each participant should bring a laptop computer on which the current version of R and RStudio are preinstalled, and on which they have the necessary permissions to install packages. See the attached software and data appendix for instructions on how to download the software and necessary data used in the class.

Schedule

Time	**Topic**
July 8th	
14:00 - 14:30	Introductions and course overview
14:30 - 15:00	Why apply automated content analysis to web and social media data? Overview of theories, methods, and data.
15:00 - 15:30	Introduction to quanteda
15:30 - 16:00	Break
16:00 - 17:00	Applying word and text metrics to online data
17:00 - 17:30	Retrieving data from Twitter with rtweet
17:30 - 18:00	Challenge 1: Calculating word and text metrics in a Twitter data set

July 9th	
14:00 - 14:30	What are sentiment analysis and dictionary analysis?
14:30 - 15:30	Applying sentiment analysis
15:30 - 16:00	Break
16:00 - 16:30	Challenge 2: Applying sentiment analysis to Twitter data
16:30 - 17:30	Applying sentiment analysis
17:30 - 18:00	Challenge 3: Applying dictionary analysis to Facebook comments
July 10th	
14:00 - 14:30	What is supervised machine learning (SML)?
14:30 - 15:30	Applying SML
15:30 - 16:00	Break
16:00 - 16:30	Challenge 4: Applying SML to online news
16:30 - 17:00	Retrieving data from MediaCloud
17:00 - 18:00	Brainstorming project ideas / practical code recipes / one-on-one feedback

References

July 8th:
Benoit, K. (2017). Getting started with quanteda. Available at https://quanteda.io/articles/quickstart.html

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit. *Digital Journalism*, 4(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598
Lazer, D., & Radford, J. (2016). Introduction to Big Data. *Annual Review of Sociology*, 43(1).

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, 346(6213), 1063-1064.

Wickham, H., & Grolemund, G. (2016). *R for Data Science*. London; New York: O'Reilly.

July 9th:
Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), 24-54.

von Nordheim, G., Boczek, K., Koppers, L., & Erdmann, E. (2018). Reuniting a divided public? Tracing the TTIP debate on Twitter and in traditional media. *International Journal of Communication*, 12, 548–569.

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

July 10th:
Boydstun, Amber E. 2013. *Making the News: Politics, the Media, and Agenda Setting*. Chicago: University of Chicago Press

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(02), 311-331.

Bio

Cornelius Puschmann is a senior researcher at the Leibniz Institute for Media Research in Hamburg where he coordinates the international research network Algorithmed Public Spheres, as well as the author of a popular German-language introduction to content analysis with R (inhaltsanalyse-mit-r.de). He has a background in communication and information science and is interested in the study of online hate speech, the role of algorithms for the selection of media content, and methodological aspects of computational social science.