YOUTUBE DATA: WHAT WE CAN COLLECT AND HOW WE CAN UTILIZE

Taehee Kim
University of Oldenburg

@ Workshop: Social Media Data University of Konstanz

TOPICS OF THIS SESSION

- Introduce recent social science research utilizing YouTube data
- What and how we can collect YouTube data?
- Introduction to YouTube APIs
- Obtain YouTube API access
- Collect YouTube Data using R package **tuber**

ACKNOWLEDGEMENT

Automatic Sampling and Analysis of YouTube Comments, 2022 GESIS Training Workshop
Julian Kohne, Johannes Breuer, M. Rohangis Mohseni

WHY YOUTUBE?

- The most viewed online video platform
- Especially for adolescents, it partly replaces TV (Defy Media 2017; Feierabend 2016)
- YouTubers are social media superstarts (Budzinski and Gaenssle 2018)
- YouTube statistics (Baertl 2018)
- How to use YouTube data for social science research (Konjin et al. 2013; Arthurs et al. 2019)

PREVIOUS STUDIES

- Homophily and radicalization
  - Radicalization (Ribeiro et al. 2020)
  - Supply of and demand for conservative political contents on YouTube (Munger and Phillips 2020)
  - How YouTube algorithms create far-right groups (Kaiser and Raucchfleisch 2020)
  - Opinion based homogeneity (Roechert et al. 2020)
- Ethnic insults in YouTube comments (Spoerlein and Schlueter 2021)
- Gender issues (Wotanis and McMillan 2014; Doering and Mohseni 2019)

HOW TO COLLECT YOUTUBE DATA?

...provides **YouTube API**.

PREPARATION

- Set up API Access: Check contents from Kohne et al. 2022 (Github): YouTube API set up instruction
- R > version 4.0
- **tuber** package in R
  - Search videos
  - Collect channel info, video info, comments, subtitles

OTHER OPTIONS

- Connecting to API without packages: link
- Other R packages
  - vosonSML, VOSONDash: focusing on comments
  - youtubecaption: focusing on captions

YOUTUBE API

- YouTube API Overview
- All possible resources
- Most utlized sources among social scientists: **search, comments, video information**
- It returns JSON formatted data
- API keys or OAuth 2.0 ?
  - API Keys: you can access to all public data
  - OAuth2.0: Token created from Client secret and Client ID, you can access to everything the user can access
  - OAuth2.0 is required to use **tuber**
- API rate limits: you have a quota of 10.000 units per day
  - Each request costs a certain amount of units
  - Rate limits
  - quota you have already used up

TUBER PACKAGE

- tuber package
  - it handles authentication, provides functions to request API calls
  - transforms JSON formatted data to R dataframe
  - but they also have some issues: do not fetch all comment threads. For more detail see their github repository issue section.
- Collecting YouTube data using **tuber** package
  - Check Kohne et al.'s slide