Can ready-made language models be used for context-specific coding?

Categorizing Twitter actors using large language models and APIs

Johanna Einsiedler and Simon Ullrich (SODAS, University of Copenhagen)

johanna.einsiedler@sodas.ku.dk; simon.ullrich@sodas.ku.dk

Over the past year, the capability and accessibility of large language models (LLMs) has massively increased. In our work, we explore how LLMs in combination with automated web search can be used to significantly simplify even complex, traditionally time intensive, actor categorization processes. In so doing, we go beyond existing approaches to actor coding which have either been fully manual or based on training or fine-tuning of context specific machine learning models (e.g. Yan et al., 2013). We argue that leveraging these recent advances for social science research increases the efficiency and reproducibility of fundamental steps in research procedures. It also enables human-machine interactions than can assist or even augment the researcher and bridge the long-standing divide between computational-quantitative research and interpretative-qualitative approaches.

We use the case of a categorization of actors engaging in Danish Twitter discourse on nuclear power to test and illustrate the affordances of LLMs to facilitate coding procedures. Categorizing social media users is a fundamental problem in social media analysis. Specifying categories of users engaging in online public spheres is also relevant beyond purely actor-centered research as it enables, for instance, the analysis of the relation between the social (*who?*) and cultural (*what?*) layers. However, when working with large social media data, research is challenged by material consisting of traces not being created for research purposes. Apart from the sheer volume of unstructured information, inconsistencies in the data present a significant obstacle. For instance, while some Twitter users provide more detailed information in profile descriptions and other meta-data, others do not.

To a limited extent, scholars addressed the problem of inconsistency in available information and have triangulated Twitter with other online sources (Burger et al. 2011). Yet, fully manual coding is still widely employed in the social sciences (Do et al. *forthcoming*), let alone in the qualitative strand. Partly, this is due to computational classification methods not being easily applicable in more qualitative methodologies. Existing classification models are usually developed deductively, focus on a single task, and do not accommodate the need of qualitative scholars for case-tailored categorizations. Also, the exclusive reliance on data found within the boundary of a single medium fails to account for the varying depth of information provided by actors.

We address these problems by integrating LLMs and API-based web searches into the qualitative coding process. We set out from an inductively developed codebook that contains nine social fields as main categories (among others: political party actors, NGOs, media actors, and energy business actors) and several subcategories to further distinguish between energy experts, political actors, and everyday users. We split up the complex classification task into multiple, carefully defined sub-tasks that can be solved computationally (see *Figure 1*). Thereby, we show how general-purpose language models can assist researchers even with complex, context-specific actor categorizations. This form of computer-assisted coding serves as a meeting point for data science and social science and bridges quantitative and qualitative method registers to advance our understanding of cultures and cultural production.

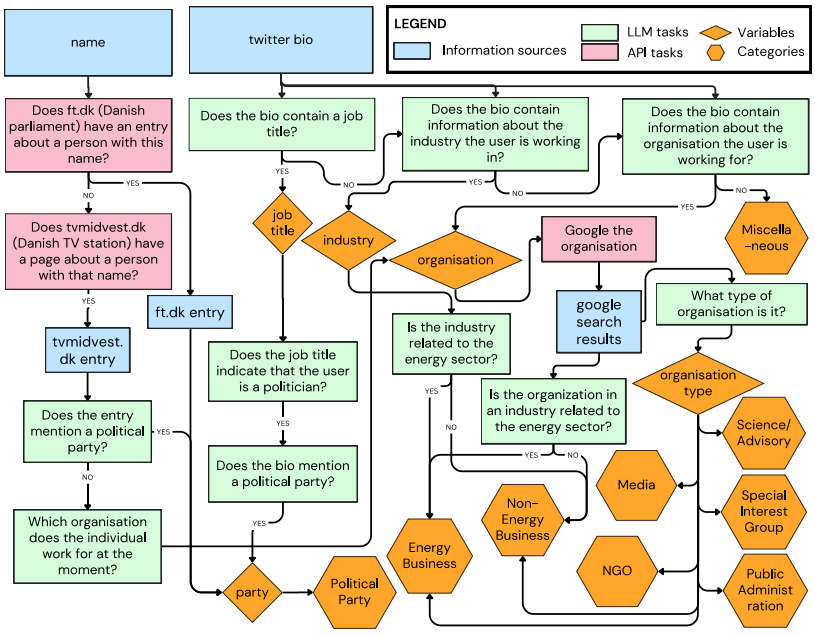

Figure 1. Workflow chart

The chart illustrates the basic workflow for the computer-assisted actor categorization. It shows the multiple steps, data sources, and computational methods applied to assign individual Twitter users to pre-defined actor categories.

Literature

Burger, John D. Henderson, John., Kim, George., Zarrella, Guido. 2011. "Discriminating gender on Twitter." In *Proceedings of the conference on empirical methods in natural language processing*, 1301–1309. Association for Computational Linguistics.

Do, Salomé; Ollion, Étienne.; Shen, Rubing. Forthcoming. "The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy." In *Sociological Methods and Research*.

Yan, Liang, Qiang Ma, and Masatoshi Yoshikawa. "Classifying Twitter Users Based on User Profile and Followers Distribution." In *Database and Expert Systems Applications*, edited by Hendrik Decker, Lenka Lhotská, Sebastian Link, Josef Basl, and A. Min Tjoa, 396–403. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013.