

Large language models to supercharge and humanities cultural analytics research

Andres Karjus

ERA Chair for Cultural Data Analytics, Tallinn University

The increasing capacities of instructable large language models (LLMs, sometimes dubbed as AI) presents an unprecedented opportunity to scale up data analytics in the humanities, cultural studies and social sciences, and to automate qualitative tasks previously typically allocated to human labor. Of particular interest to the humanities and social sciences is the capacity to use them as zero-shot classifiers and inference engines (cf. Ziems et al 2023). While classifying texts or images for various properties has been available for a while in the form of supervised learning, the necessity to train such models (or even tune pretrained models) on sufficiently large sets of labeled examples has arguably hampered their adoption for research beyond generic tasks e.g. sentiment analysis. Approaches like word or sentence embeddings and topic modeling allow for explorative approaches but are typically laborious to interpret and difficult to use for confirmatory inference.

We report on eleven experiments of applying LLMs in zero-shot classification and reasoning scenarios derived from or replicating existing research. These cover various practical, exploratory and confirmatory tasks, including linguistic feature analysis, political stance detection in a lower-resource language, text reuse detection, literary genre inference, historical news topics prediction, wreckage cause inference from maritime databases, social network inference from literary texts (see Figure 1), novel word sense inference, semantic change and divergence quantification, contextual missing data imputation, and relevancy filtering of examples mined from digitized newspapers. Given the prevalence of English focus in the emerging LMM literature, most of these examples deal with harder scenarios, including data in languages other than English, or historical texts prone OCR errors.

Regardless, we report fairly high to almost perfect agreement with human annotations across the board (using OpenAI models gpt-3.5-turbo and gpt-4 as example LLMs). LLM outputs naturally contain errors (as does human annotation), but the error rate can and should be included in subsequent statistical modeling. We also stress the importance of task-specific test sets and caution against relying only on general model evaluations (including those reported in this contribution), even if thematically related.

In summary, we argue that in all but the most difficult tasks requiring expert knowledge, LMMs can serve as a viable analytic alternative to human annotations and can be used in mixed methods frameworks that involve quantification and statistical modeling of qualitatively annotated data (e.g. usage feature analysis in linguistics). Unrestricted by the limitations of human analytical capacity, this development promises to scale up research in the humanities and social sciences to handle much larger volumes of data.

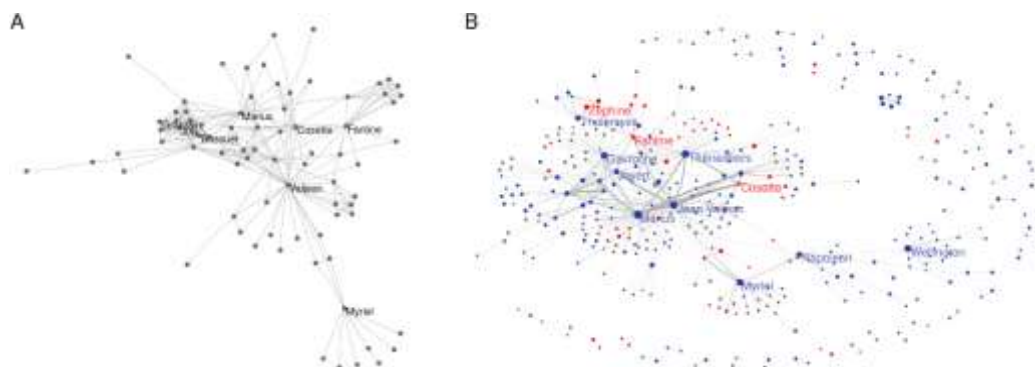


Figure 1. The classic Les Misérables network (A) often used as an example in network science, and a network directly inferred from the same book (B) using an LLM instructed to extract interacting characters and output them in a systematic machine-readable format, additionally tagged for gender (red = women, blue = men).

References

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. 'Can Large Language Models Transform Computational Social Science?' arXiv. <https://doi.org/10.48550/arXiv.2305.03514>.