Large Language Models in the Humanities: magic shortcut or just beating about the bots? #nocode #alldata

Gunther Martens, Ghent University

Lore De Greve, Ghent University

Following up on our earlier contributions on doing Aspect-Category-Opinion-Sentiment Quadruple Extraction, in this talk, we aim to weigh some of the tried and trusted methods of doing NLP against the new kid on the block, namely LLM (Large Language Models) trained for various purposes and languages. Few-shot and zero-shot approaches may seem adventurous, and they indeed may not live up to the state-of-the-art achieved with regard to our specific use-case, i.e. aspect-based sentiment analyses (ABSA) in specific domains such as laptops and restaurants. But what if e.g. your target vocabulary is out of domain (OOD) by principle? What if your data is largely unstructured full-text? Compared to the time-consuming task of generating annotations, using GPT-3 turns out to yield surprisingly good results for a variety of tasks like NER and ABSA. We will show how to "escape" the chatbox and feed data in bulk to the model. We aim to discuss whether this is a viable solution, especially for the aspiring DH researcher with limited programming skills, limited GPU access, and/or without persistent support from teams dedicated to humanities-style qualitative research questions. We will present some of the tasks that have become possible over the course of conducting our study of multilingual sentiment in a specific multimodal setting, such as: automatic transcription of oral sources, alignment of oral and written data, speaker diarization, automatic generation of training data with annotations, etc. We will illustrate the major advantages of API-based solutions for classification, such as the small amount of training data needed to fine-tune the model, and the ease of access. All in all, this type of modular pipeline is a less prohibitive pathway for aspiring researchers, compared to setting up a demanding and complicated environment that is typically required in traditional NLP methods. In addition, allowing people to "chat with your data" will inevitably become a new way of disseminating research results in a non-directive way, also in domains where access to the actual artefacts might be marred by copyright. This workshop is not about presenting results or final conclusions, but aims to provide an opportunity to discuss what some consider to be a watershed moment, while others remain sceptical.