Workshop report: Harnessing Generative AI for Data-Driven Journalism (2 hours)

Led by: Borislav Vukojevic, Senior Teaching Assistant and AI consultant, University of Banja Luka, Bosnia and Herzegovina

Objectives:

1. Understand the capabilities and differences of various AI tools: ChatGPT Advanced Analyst, Claude AI Analyst, Julius AI, and Google AI Studio.
2. Learn how to integrate these tools into the data analysis process.
3. Explore real-world applications of AI in investigative journalism.

The workshop commenced with addressing several key aspects of and misconceptions around working with Gen-AI in journalism:

- For any work with AI, but particularly for data analysis, one needs to fully understand how individual AI tool work, what is their capacity, the analysis models they use, and what are the differences/strengths of individual tools, so as to utilize the best one for a given task
- **A major problem with how most people use AI is that they treat it like an "answering machine" (aka search engine)**, where in reality, as most AI tools are still predominantly LLMs (large language models), they were never meant for this. They are language processors, focused on statistical relations between words. Therefore, they lack the ability of "abstraction and imagination" as well as the real world context knowledge humans have. Thus, they cannot be reliably used for answering complex questions. Some exception here is e.g. the *Perplexity* tool, a hybrid between LLM and a search engine.
- Therefore, when working with AI tools, it's always good to essentially treat them like a person dependent on being given the full context necessary for processing the task at hand.
- **Capacity limitations**: the most important aspect of any AI tool is its "context window", the processing capacity the given tool has. In other words, the number of parameters a LLM can support in one conversation. Different tools (and their free/paid versions) have different sizes of context windows. For instance, Chat GPT has around 35 thousand tokens (characters)[1], whereas Google's Gemini 1.5 Pro has up to 2 million tokens. The context window directly influences how much data can you upload into the system for analysis. Crucially, not all tools tell you if their context window is overstretched (i.e. too many data was uploaded), and wherever this happens, the results you get stop being reliable or become completely flawed, as the tool "makes up" for its insufficient capacity

[1] As tested by the workshop leader, otherwise Chat GPT advertises thousands more.

in different manners, like data repetition or omission. Chat GPT especially is known not to disclose this information, unlike Gemini or Claude.
- Transparency of methods used: every AI tool has, apart from capacity, also different default methodology. This will also differ based on the free/paid version of your account. If you wish the tool to conduct an advanced or a particular kind of analysis, you have to clearly instruct it to do so.

➔ These misconceptions feed into many people's reservations about using Gen-AI for their work, but particularly for data and investigative journalists, if used properly, these tools can immensely increase efficiency and save time. Our misunderstanding of the individual tools' working and capacities yields wrong results, and people get too easily discouraged. "Waiting several years" until the technology is more advanced enough was not recommended by Mr Vukojevic, as that attitude will put you at a comparative disadvantage threatening your own professional life. What is more, we are at present at the wide experimental stage of using Gen-AI. This is the perfect time to experiment, share our experiences with the tools with our colleagues, so that we can both develop the tools into more quality aids, and mutually learn about using them more effectively.

Several key points to better navigate the AI tools for data analysis:

- The "Prime, Prompt, Polish" model is worth following.
- **Priming** is the essential part which will wholly define the success of your analysis. It is vital you give the tool **clear examples** of how you believe the analysis should be done (e.g. from previous investigations). If you give your own examples, it will mimic your manner of analysis better. Mr Vukojevic recommended **using XML (Extensible Markup Language) examples** for this stage, because they're the best at explaining to the AI tool/LLM which part of the data submitted for analysis is important, what it is and how should the tool treat it. It's essentially like bolding a given information in a text. XML codes are also easier for AI tools to learn. You can also predefine the codes in the "custom" instructions of the given AI tool and even give shortcuts for them.
- Similarly, if you wish the tool to incorporate further information or verify it against something on the internet (different tools have different search options for this), you need to give clear examples of websites/domains the tool could get the info from. If you wish for the tool to draw on many sites, best results are achieved if given 8 examples of such sites.
- In priming you should get the tool as much **context** relevant to the analysis as possible. This, too, can be done through examples (e.g. various conventions on human rights for giving gender discrimination context)
- **Prompting is the stage where you can customize the knowledge base the tool has for the analysis.** Ideally, create your own set of prompts, so that you can replicate them over subsequent analyses. As a rule of thumb, for quality work with Gen-AI, everything

you teach the tool should be related to you – your examples, your data your methodology, your way of writing/thinking, your prompts. This will both yield better results and serve for their better verification as you will be able to spot any potential issues more easily. Buying prefabricated "mega prompts" is a waste of money for serious professionals.

- **Always use English language as the input language for priming**, as LLMs were taught with this language and can thus learn further with it more easily. Use other, native languages for prompting as the next step.
- When choosing AI tools for various tasks, one important variable to keep in mind is their "**LLM temperature**" capacity. This parameter controls the randomness of text that is generated by LLMs, in other words, decides the „creativity" and approximation of human thinking in the analysis. Zero temperature value connects only the most predictable values and thus gives only the „safest" results. This is good for data analysis, but for other tasks like writing policy briefs, articles, essays etc., tools with higher temperature give better results.

➔ The workshop then moved to **comparing several of the most efficient AI tools**, exploring their strengths and weaknesses, as well as different ways of using them. Currently, it's recommendable to use several tools/LLMs for one investigation. Draw on their individual strengths, compare the results, modify as necessary. There is no "one best tool" for everything.

1) Chat GPT Advanced Analyst

- When using the free version, the context window is 8 thousand tokens (but it will not indicate so), so it's very limited. Neither will Chat GPT let you choose the methodology of analysis; it will be assigned to you based on your prompt. So, for proper analysis you do need to have a paid account.
- It can be used in two standard ways:
    A) prime conversation -> upload data set -> prompt for insights -> modify as necessary
    B) use custom GPTs (better and replicable priming, but severe drawback is no info about the context window even in the paid version)
- After you upload a data set, Chat GPT gives you a set of possible prompts and questions – these will yield good results because the machine offers you information it can reliably give you. But for using your own questions, which are different and quite likely more complex, the quality of output diminishes.
- Mr. Vukojevic then showed the difference between asking and not asking for the advanced model of analysis to be utilized. For small datasets, Chat GPT works fine, esp. with the advanced methodology, but is not recommended for big data.

2) Claude AI Analyst

- According to Mr. Vukojevic, at present the most reliable tool (again, in the paid version)
- It does show a system error if you overstep the context window
- It has the "Projects" function, which corresponds to custom GPTs but enhanced by that it tells you how much of the context window you use -> this option is ideal for priming your own **knowledge base** for working with the tool, without the need to reupload it for every conversation and teach the system anew. It will also get increasingly better and more tailored to your individual needs with every new prompt and your feedback.
- It also newly has a function of incorporating your "style" which allows it to produce outputs in a way you write and think, drawing on insights you have given it from your own work (at the priming stage).
- It also has the benefit that unlike GPT it can render codes, not just write them, and thus it can give you e.g. interactive visualisations of the data or even create a website with the results directly inside the tool.

3) Julius AI

- Combines Python (a programming language used in statistics), Excel and ChatGPT -> a tool fine-tuned for statistical analysis, which makes if perfect for more advanced data analysis.
- Enables choice of which LLM will be used for the analysis
- Also good for data visualization and clearing the data based on prompts already inside the tool, also for running statistical tests
- So far the best for setting up the aforementioned workflow by priming and subsequent prompt modification

4) Google AI Studio

- Best for very large data sets
- Enables choice of LLM for analysis
- Context window of up to 2 million tokens, which is the largest on the market (can fit all data about an average medium-sized company into a single conversation)

Concluding points:

- For those working with sensitive data (like investigative outlets often do), always work only with an internal locally uploaded version of the tools, do not use the generally available front desk option on the web. This is safer and partially also works around the data privacy concerns
- It is recommended, for any media outlet, to pick AI tool best suited for the needs of the outlet or individual departments, create a customized set-up within it with extra careful priming for best results and adaptability of the given medium. Subsequently, all relevant staff should be trained in using it and get acquainted with established internal rules of use. It's vital to understand this is not just another purchased ready to use license tool, it needs to be further developed for it to be meaningful and financially viable.
- Open AI's o 1 model is fist from a new range of tools which will make another revolution, since it's designed to be not a basic LLM, but an actual "thinking model" capable of showing the users its "chain of thought". This means it will in essence be doing the priming for the user, and the user will also be able to double-check the rationale behind the tool's analysis and change any part of the chain which may be flawed. This tool is ideal for scenarios building and predictions, adding another layer of added value for data analysis and journalists.
- For good analysis, it's necessary to use the paid versions of the tools. These can get very expensive, but there are smaller tools accessible even for smaller outlets, and more importantly, as the market will be expanding, it is expected the license prices will be decreasing in the future.

Further research:

- https://onlinejournalismblog.com/2024/06/06/ai-in-investigative-journalism-mapping-the-field/
- https://reutersinstitute.politics.ox.ac.uk/news/i-created-ai-tool-help-investigative-journalists-find-stories-audit-reports-heres-how-i-did-it
- https://gijn.org/stories/new-ai-large-language-model-tools-journalists/
- https://pressgazette.co.uk/platforms/how-ai-could-save-investigative-journalists-time-and-test-their-hunches/