Mining arguments

Workshop Week & Pitch Night (21st - 25th June 2023)

Project report in *SustAInability* at the 'Hochschule für Politik München' of the Technical University of Munich

Supervisor   Helene v. Schwichow & Charlotte Böhm

Tech founder Vinzent Ellissen

Authors   Karolina Senoner

Annika Markoff

Onur Alpaslan

Nicholas Menzel

Submitted   Munich, 02.07.2023

Table of contents

1. Introduction

1.1. Background

This report sums up the workshop week and pitch night of our seminar SustAInability under the supervision of Helene v. Schwichow (TUM) and Charlotte Böhm (HM). The project took place at the TUM Think Tank in the "Hochschule für Politik München" (Richard-Wagner-Street 1, 80333 Munich) from the 21st to 25th of June 2023. Before the week began, students chose between six different challenges located at the interface of Sustainability and AI:

- MI4People: Arne Hartz and Paul Springer, Finding Garbage in the Ocean: Can you beat the experts?
- AI Transforms the Fashion Industry for Circularity: Carina Legl
- Mining Arguments: Vinzent Ellissen
- Does the EU AI Act Promote Socially Sustainable AI? Felix Rank -EME
- AI Model Cards – a new concept to standardize environmental reporting for AI models Johannes Kirnberger, OECD
- "Unlock Circularity and Avoid Waste" with Stefan Hauser

1.2. The challenge

A total of eight students chose to take part in the third challenge with Vinzent Ellissen – mining arguments. Divided into two different groups, both groups tackled the same challenge, bringing different approaches to light.

The "Mining Arguments" challenge asked for a solution to problems concerning misinformation and weakened democracy. The goal of the challenge was to figure out a way to utilize LLM's (large language models) to extract arguments from given texts, with the ultimate goal of improving ease of access to information for consumers and policymakers. By making it easier to process larger amounts of articles, a user would be able to reach informed decisions more easily. The challenge, as our group tackled it, had two aspects: first, understand the current research on argument extraction. Second, create prompts and flows of prompting (e.g. Tree of Thought) to yield the best results. The goal at the end of this challenge was to be able to apply our prompt or series of prompts to 1-3 real life examples that we were most passionate about, preferably with strong ties to sustainability.

2. The project week

2.1. The challenge giver

During the challenge we were well supported by Vinzent Ellissen, who is a tech-founder, Managing Director of Undiced GmbH which currently focuses on making secure online publication easier and exploring several approaches for applying Artificial Intelligence in communication. He volunteered in the steering committee and as chair of the board of the young group of the Bavarian Economic Forum (Wirtschaftsbeirat Bayern e.V.) which is an influential and politically independent professional association that fosters the dialogue between politics, economics, science and citizens.

2.2. The support

During the week we got very useful and insightful feedback and suggestions from our supervisors Helene v. Schwichow and Charlotte Böhm, as well as comments on our interim presentation from Prof. Stefan Wurster, Prof. Gudrun Socher and Maximilian Dauner (PhD Candidate at HM). We met twice for one hour with our challenge giver Vinzent Ellissen and we were able to ask and communicate with him via mail or WhatsApp during the project.
At first, however, it was difficult for us to understand what was specifically required. It was just after the first meeting with Vinzent Ellissen that we cleared our minds.

2.3. The goal

- Goal: Help create informed citizens and policymakers, strengthening their ability to make decisions

- Task: Given an article, get a large language model (LLM) to find an author's main arguments, the basis for those arguments, and the source of the information

- End product: A prompt, or series of prompts, that guide the LLM to providing the necessary info

2.4. Our procedure

Simply described, what we did the whole week was to try out, in a systematic way, different prompts in ChatGPT. Given an article, we used ChatGPT as a large language model (LLM) to find the author's main arguments, the basis for those arguments, and the source of the information. We have not limited our search to specific articles but tried out the prompts on different texts such as newspaper articles as well as research papers.

We generated multiple initial prompts using research about prompt generation and prompts generated by GTP 3.5 and GPT 4.0. These prompts included definitions of the ChatGPT persona, Tree-of-Thought prompt chains, setting of temperature and frequency penalty, full examples and/or example output structures. Using articles that we had manually analyzed for all their arguments, each of these prompts was evaluated on an article and documented in a spreadsheet along with the source article and the output from ChatGPT. We also included general feedback about the prompt and what insights we gained by trying out this prompt. These prompts were then further updated and documented. In this step we ended up with 36 different prompts.
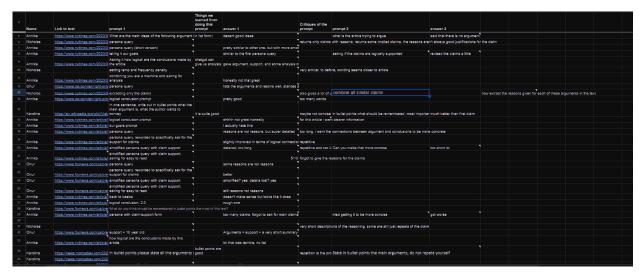

Figure 1 List of prompts and outputs

Using the articles and the generated arguments we evaluated the prompt on different criteria:

- Presentation

- Logical consistency

- Length

- Ease of understanding

- Repetition

- Objectivity

Original Prompt Base	New version of prompt	Gives information in a list	Supports logically support the claim	not too long, not too short	Easy to understand	Not too much repetition	Is objective
Persona query							
	persona query (short version)	10	8	6	5	3	3
	persona query, reworded to specifically ask for the support for claims	10	9	9	9	4	9
	persona query D6	10	5	5	8	4	7
	What do you think should be rem	10	9	9	7	5	10
	setting temp and frequency penalty	8	5	6	6	7	4
	simplified persona query with claim support	10	9	2	9	9	10
	simplified persona query with claim support, asking for easy to read	10	3	7	7	3	10
	persona query	10	3	3	7	3	10
	extracting only the claims						
	persona query, reworded to specifically ask for the support for claims	10	8	8	8	3	9
	support + 10 year old	10	9	5	8	9	9
	simplified persona query with claim support	10	3	3	6	3	9
	list, simple, supporting reasoning	10	6	9	8	8	8
	simplified persona query with claim support, asking for easy to read	5	9	6	7	5	10
	persona with claim/support form						
Logical conclusion base		10	9	4	4	4	9
	logical conclusion prompt	8	9	5	5	7	7
	how logical are the conclusions made by this article						
telling it our goals		10	6	7	5	6	7
	combining you are a machine and asking for						

Figure 2 Prompt evaluation

Using this evaluation, we found the different strengths of different prompt implementations and used these to generate three final prompts to use for further testing.

These prompts were then tested on multiple different articles to try and improve them and make them more consistent. The workflow for testing and improving the consistency across multiple articles is shown below:

Figure 3 Consistency testing and further improvement

This finally led us to our final prompt.

2.5. Final product

Our final edition of the prompt reads as follows:

Disregard any previous instructions.

Persona: You are a professional analyst working for a reputable research organization. Your role is to analyze and extract key arguments from various articles objectively and present them in an emotionally neutral tone. Your goal is to provide an unbiased understanding of all of the main claims and the support the author uses for those claims presented in the article.

The format should be as follows:

Claim 1: (Insert the first key claim here. Make sure the reader can grasp the context. Avoid using demonstrative pronouns.)

Basis: List here all the support the author provides that directly supports the claim the author has made. Provide citations where relevant. Use brief bullet points. If the author has provided no convincing evidence, fact, logic or opinion, write, "Insufficient support was given". If there was no citation given, show the findings and avoid writing "No citation specified for this claim."

Coherence & Logical Structure: Evaluate the claims. The criteria for evaluation should take into account the logical structure, coherence and consistency of the argument, regardless of the stance or tone. Fallacies, ungrounded arguments, and unsupported claims should decrease the

strength of the claim. Show the sources and citations used as well as the amount of uncited arguments. Keep it brief and use bullet points.

Claim 2: (Use the same structure as before and create as many claims needed)
Basis: (Use the same structure as before)

Claim 3: (same)
Basis: (same)

Create as many claims as there are in the input text.

Here is the article:

This prompt implements the features clarified in the development phase of the project. The aspects we determined to be important while testing prompts were:
1. Priming ChatGPT to relay the answers in an objective tone by introducing a "persona"
    a. "You are a professional analyst working for a reputable research organization."
2. Specifying a format, ensuring consistency
    a. "Claim 1: (Insert the first key claim here."
    b. "Basis: List here all the support the author provides"
3. Using keywords like "brief" to ensure concision
    a. "Use brief bullet points."
4. Paying attention to word choice, especially the use of words like "basis" and "support" instead of "evidence" to help cement logical connections
5. Asking ChatGPT to evaluate the claims for logical fallacies
    a. "Fallacies, ungrounded arguments, and unsupported claims should decrease the strength of the claim."
6. Specifying that the structure should continue throughout the prompt
    a. "Claim 2: (Use the same structure as before and create as many claims needed)
        Basis: (Use the same structure as before)"

These specifics in word choice and structure responded to the challenges discovered in product development, the main one being the inherent inconsistency of ChatGPT. Because it is a generative model, inputting the same prompt into different articles produces radically different outputs. For example, without specifying a structure, ChatGPT decides on a format best suited to each

article, leading to a wide range of formats. With the same prompt, ChatGPT might use a bullet pointed list for one article, and paragraphs for a different article, giving no indication as to which might be used for which. Additionally, ChatGPT is susceptible to small word differences like the semantic disparity between "fact" and "support" (one being simply information, and the other being inherently connected to a claim), meaning that many rounds of experimentation were needed to land on the final wording.

Part of the challenge, in addition to leading ChatGPT to producing the required information, lay in finding a wording that produced very similar results every time, in order to be able to easily compare the claims of different authors and absorb more information about a topic.

While our final prompt is far from being tested on every type of article and stance, we managed to narrow the output types and formats into a small enough box to process a wide range of articles.

3. Limitations

What would society look like if products such as ours were to be used widely? That question remains somewhat unclear. More academic research is needed to determine how different information channels affect social habits, for instance. A not so unreasonable argument is 'Would not the use of such software make people lazier?'. The amount of time needed to access information has drastically reduced in the 21st century. Although our end product aims to tackle misinformation, it remains unclear whether it would make the end-user lazier. On the one hand, the end-user is left with more and accurate information; on the other hand, they may not check the original sources themselves and make up their own interpretation of what they have read. In this sense, in the continuous use of information-accessing tools such as the one we propose, the end-user is limited to the meaning provided by our prompt. This leaves room for potentially biased perspectives.

How our product would be applied to works of art is a whole other topic for discussion. How do we interpret the meaning of the arguments we mine? Let us take the example of 'Animal Farm' by George Orwell. The book relies heavily on metaphors to create a satire of the Soviet Union. Not in one page, the words 'Soviet' or 'Lenin' appear, yet we all know that the book uses specific metaphors for specific events and people. Therefore, would our prompt yield results based on what the text literally reflects, or what the text actually means through the use of a certain tone and narrative? This raises a second important question and that is 'What is the actual meaning of any text?' The highly subjective nature of language currently restricts our interaction with AI. Therefore, would works of art, or metaphorical texts (such as the Holy books) be exempt from argument mining? There is room to debate for philosophers and linguistic experts.

Lastly, the prompt is formulated in a rather long way. The time concerns restricted the efficiency and effectiveness of the generated final prompt. With more time, there is room for improvement for a shorter and more effective prompt.

4. Future outlook

We tried to keep our criteria as objective as possible by using terminology such as 'logical structure', or 'coherence'. However, the meaning attributed to these specific words may differ from one to another as well. In this sense, perhaps a criterion on objectivity. The criterion for objectivity is subjective. Therefore, a consensus on the requirements of objectivity could be deliberated upon. The answer to 'What to mine?' gets increasingly hard as the input text gets more complex and abstract. More research on the field is necessary. Lastly, perhaps we need to know more about the impacts of such a system on human behavior. Would it make people lazier is research topic that could be conducted by disciplines such as behavioral psychology for example.

Although its potential weaknesses exist, we believe that the benefits of the argument are also substantial. Philosophical debates on objectivity aside, in the short run, argument mining could drastically reduce misinformation, and contribute to a more educated society with a consolidated 'information literacy'. This way, more informed citizens could be believed to take part in democratic decision-making processes, which would potentially yield healthier outcomes.

5. List of figures