Report on the outcomes of a Short-Term Scientific Mission[1]

Action number: CA21129

Grantee name: Dren Gërguri

Details of the STSM

Title: Testing ChatGPT ability to express opinions: A study of how events are perceived and communicated in the Albanian language

Start and end date: 20/07/2023 to 01/08/2023

Description of the work carried out during the STSM

The purpose of our research is to evaluate if ChatGPT is able to express thoughts in the Albanian language and comprehend how it frames certain occurrences. The first part was to discuss theoretical challenges regarding large language models, such as ChatGPT. The second stage was to develop a methodology to assess the factual accuracy of ChatGPT's responses. Compare the factual accuracy of its statements with credible sources to determine the extent of misinformation. The third part was to collect a varied dataset of event triggers in Kosovo and Albania, choosing a variety of events that cover a variety of themes and offer detailed descriptions or summaries of each event. The next step was to gather and capture the ChatGPT answers for each event prompt in the dataset. Ensure that the event prompt, associated produced answer, and any extra relevant information are properly documented. The last part of the project was to assess response quality (coherence, factual correctness, etc.) as well as compare and contrast replies in order to find trends and variances in how ChatGPT generated responses.

As a starting point in this analysis, a database of 80 communications with ChatGPT on various issues has been created. The communication is carried out from different IPs so that it is not affected by the algorithms of one user. After the communication, the next step was to analyze the responses of ChatGPT, in order to see if there are Opinionated AI or statements based on the trained data.

Description of the STSM main achievements and planned follow-up activities

The objectives of the STSM were achieved in several stages. The first stage included the preparation of the study design. Together with the host of the STSM, we discussed the theoretical framework, methods, and data collection principles. The method chosen for the study was a combination of quantitative and qualitative methods, including content analysis to identify patterns in how it frames and presents texts, and also assessing the factual accuracy of ChatGPT's responses and the contextual factors that influence ChatGPT's opinion presentation, including the input prompts, user demographics, and prevalent discourse

[1] This report is submitted by the grantee to the Action MC for approval and for claiming payment of the awarded grant. The Grant Awarding Coordinator coordinates the evaluation of this report on behalf of the Action MC and instructs the GH for payment of the Grant.

in its training data. The data collection constituted a very important step in the study as we had to choose a variety of events that covered a variety of themes. After creating the dataset, we began interacting with ChatGPT and generating replies to each event prompt in the dataset. On the theoretical part, we decided to use the framework of Jungherr and Schroeder that allows for empirical examination of AI's structural impact on the public arena. The AI is used in structures to carry out numerous functions in three main categories: shaping information and behavior; generating content; and communicating. Our main focus is in the category of generating content. As Jungherr and Schroder emphasize, autonomously generated content can offer trustworthy information to the public arena, but AI can also overwhelm the public with material that leads to the decline of information quality. Our paper that we are working on will deal with this matter, how much AI-generated content contributes to the deterioration of information quality. Based on the preliminary results, a focus group experiment also will be prepared and conducted.

Large language models may alter how we develop ideas and influence one another in regular conversation. Large language models that generate language similar to that of humans are now becoming more common, which means that interactions with technology may affect not just behavior but also opinions. These models may disorient their users if they provide some inaccurate information frequently. Besides that, there is a concern that Large language models could replace writer or could be an opinion leader. Here we were interested in how ChatGPT could impact people's opinions, for instance, the latent persuasion (when language models produce some opinions more often than others). Previous studies offer examples of how interaction with others may change our thoughts and attitudes. Our research idea is to focus also on how user interactions, including the framing of questions and requests, affect the tone and stance of AI-generated opinionated communication. We discussed to use for this analysis of the three potential trajectories of influence inspired by social influence theory: informational influence, normative influence, and behavioral influence.

During the STSM, we discussed about the appropriate method for collecting a diverse and representative dataset of AI-generated text. We decided to conduct the experiment using different IP addresses while we communicated with the ChatGPT. Different questions were directed to the chatbot at separate times. For this experiment, we created a list of questions related to important events and personalities for Albanians, such as Kosovo's independence, Kosovo's flag, Albania's flag, Ibrahim Rugova, Mother Teresa, NATO intervention during the Kosovo War, etc.

Our conclusion is that an opinion to be "opinionated AI" requires a system capable of generating text that expresses perspectives and subjectivity, rather than just factual or neutral statements. From our analysis of "ChatGPT experiment", we conclude that in the actual stage of its evolution, there is still no opinionated AI. The "experiment" conducted during the STSM resulted in 35% of cases where ChatGPT offered completely or partially incorrect information. However, none of them were opinions, but were statements based on the data it was trained on. There was no stance on controversial issues, for instance when asking about NATO intervention in the Kosovo War, ChatGPT summarize two different narratives regarding that, but it is not being able to do a value judgement or attitude regarding the issue. For the measurements of an opinionated AI, we thought for few questions, related to the origin of opinions (whether opinions are specified by users or derived by the system?), to transparency (how clearly the system will reveal the origin and subjectivity of generated opinions?), and to capability (whether the system will be able to connect opinions to facts and weight evidence?). The future collaboration with the host of the STSM is to complete the article and submit it for publication during next year. For the future, our plan is to do same or similar experiments to other large language models.