Technical Specification Double-side Page

1. **TECHNICAL SCOPE:** Summarize the mock-up devised during the EXPLORE phase: how have you addressed the challenge/Theme Challenges and tackled with its requirements and data. Include a diagram.

Our current mock-up focuses on the identification of additional detail levels and feedback loop for both writer and readers. The supplied data, as JSON, which contains various relevant fields for determining topic and article sentiment needs to be first parsed into individual sentences. These sentences are extracted in Dutch (or in English) and each subsequent action will take place on translated sentence structures. This matrix table is then subdivided, and each word is given the right location identifier within each translated sentence and the original location.

User control and writer feedback are integral to the theme's challenges, so we've based our mock-up around systems that provide integral information for the user, but the visual writer feedback allows the writer to influence the decisions of the AI and optimize their article for summarization. This creates an ideal feedback loop between writer and the AI algorithm. Another unique aspect is the relative customization of the duration and detail level of a summarization, since we have various parameters controlling the extractive text results, we need a way to communicate this effectively to the user.

Similarly, if we replace the traditional time-only summarization selection with a more contextual option we can also benefit from the user interaction. For this, as demonstrated in the mock-up, the user can select not only time but also generated clauses about the summary. For example: "informal", "to the point", "bare-essentials", "shortened". At the end of reading the reader is supplied with options to receive additional bits of information, the user can continue receiving more detail (additional extracted sentences) and is asked to 'rate' the result of these extractions.

2. **ALGORITHMS, TOOLS AND CONCLUSIONS:** Detail the algorithms and tools identified to accomplish the challenge/Theme Challenges. Show clear understanding of the used REACH dataset/s and addressed challenge/Theme Challenges.

Natural Language Parsers for Node.js for the tagging of titles and identification of content that is not paragraph texts. The used dataset is supplied as JSON with additional information to determine the topic of the article, but formatting data is not stored. This requires pre-processing (and user-corrections) in case erroneously titles get left out, or other information is present (such as advertisements, or unrelated snippets).

We have also determined that using the OpenAI GPT-3 base models to create and generate training data is already quite successful, this data can then be imported in AWS Comprehend and other more simple language models such as BERT, the use of AI to generate training data which enables the vast scaling of data required.

Since applying we've decided to switch to a faster development platform than Angular, which is Sveltekit. This full-stack framework offers the ability to do server side rendering giving much more control over our data and UI/UX experience should be smoother.

3. **SCALABILITY AND FLEXIBILITY OF THE SOLUTION:** Discuss whether the solution can truly cope with humongous and increasing datasets and how flexible it is to adapt to other related domains

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

Since the content in our Dataset that we're working with is not structured we are creating a system that in the end will work with any form of text, whether its paragraphs and titles or just plain text. It's likely that we're able to define edge use-cases where additional requirements and agreements on the format of the supplied text can be made which improves the algorithms capabilities because we're able to feed it more contextual data (e.g. this sentence is a title, all titles are important).

There is also the scaling for educational purposes and the subsequent verification of knowledge, such as summarizing an educational article and then questioning the readers comprehension, retention and base knowledge. The fact that we're relying on feedback loops to create both AI driven loops as well as user driven corrections, annotations and interactions creates a scalable system for vast amounts of data. Since the 'required' effort is distributed over many people this ensures that there is more benefit for the writer/reader than the cost that it takes them to annotate, correct or supply additional information or decide on the length/version of the summary that they want to read.

4. **DATA GOVERNANCE AND LEGAL COMPLIANCE:** Describe the security level of the proposed solution, i.e. how authentication, authorization policies, encryption or other approaches are used to keep data secure. Explain how will be compliant with the current data legislations concerning security and privacy (e.g. GDPR).

Users both annotate, read and create new data but can also supply their own custom information to use as an additional library. This data of course should be kept proprietary which means for example the hashing of any personal names as well as keeping the user data secret and custom for the user. Should the user decide to share their data with other parties they should be able to share within the system the secure metadata connected to the articles, which establishes a Digital Value Chain where end-users can safely share data with another even without disclosing such data. Additionally, the requirements are that there cannot be direct data processing for the front-end user that can be lead back to this user unless they actively choose to do so when give the option.

At the same time these functions also are relevant when parsing articles and creating part of the open dataset because following the GDPR, we cannot republish any source material that can be traced to the original create unless we have their permission.

5. **QUALITY ASSURANCE AND RISK MANAGEMENT:** Describe the quality process planed for the final product. Technologically, which are the potential risks in all the phases of the project (design of the solution, development, testing, deployment…) and indicate mitigation plans to still fulfil the challenge/Theme Challenges and data provider requirements.

During the design stage accurately executing the cost prediction is a potential technological risk as well as sourcing the required interaction from users/writers for the feedback loop.During the development stage the scope could change depending on the cost prediction which in turn will lead to a create a rule-based/AI hybrid system that generates the ML training data. Developing effective and clear UI/Messaging that can accurately convey the different parts of the algorithm. In the case of high costs for the most effective models and the resulting creation of more custom models Testing: Version assessments and version scoring can be a difficult challenge, especially when the models can change rapidly. It's easy to drift away from the optimal path since it takes more people to verify the effectiveness of each subsequent version due to the data-size increases, as well as increasing the required annotation effort to verify the results of the AI which is necessary for the testing stage.Deployment: Several of the pipelines are pre-created but others require live input and it could be challenging to create the real-time data deployment on a larger scale. Real-time feedback for the writer is key to the success of the project and is an integral part of our solution.

Potential mitigation plans for technical risks as described above:

- Training more custom AI models
- Directly from the beginning interact with the target audience to gauge their reaction and requirements.
- Validate the technical requirements that have the largest implication as soon as possible and consistently.
- Procurement of other interested parties/content/data providers in case we're required to spend more time on annotation.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981