Technical Specification Double-side Page

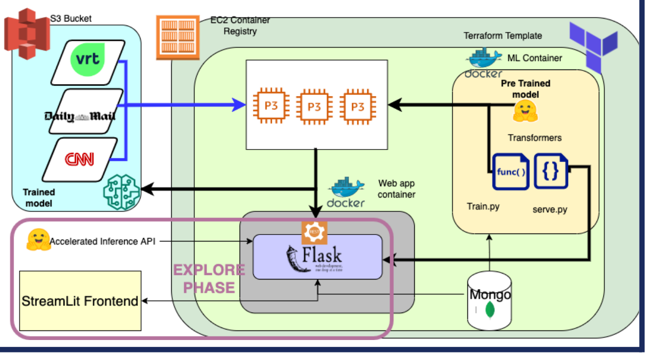
1. **TECHNICAL SCOPE:** Summarise the mock-up devised during the EXPLORE phase: how have you addressed the challenge/Theme Challenges and tackled with its requirements and data. Include a diagram.

During the **EXPLORE** phase, our **objectives** were to a) prepare a development pipeline and b) create an MVP that provides a user interface that calls the developed summarisation API.
The needs and requirements of VRT were to be able to **create multiple variations of summaries**, where the user also control the **length of each summary**. They also wanted to be able to quickly edit any summaries, and then use them as training data to further improve the model. In more **concrete deliverables for the EXPLORE PHASE,** we delivered the purple part of the diagram below:

1. Developed **helper functions with metrics and evaluation strategies**, to select the best summaries
2. **Deployed pre-trained models exposed through HuggingFace accelerated inference,** trained on news articles in Dutch, and **built API endpoints (backend)**
3. Created a **frontend Application** where we can input long text and generate various summaries

In the **EXPERIMENT** phase, we will extend the summarization problem pipeline through the following steps:
1. We will finetune the language models with VRTs data to generate summaries with the tone and style of VRT.
2. Our summarisation model will then ingest new news texts to generate a number of candidate summarizations.
3. A set of metrics rank the order of the summarizations (in terms of natural language, factual correctness, length, etc)
4. Only the best examples are served to VRT journalists to review, modify and eventually display on the website.
5. These summaries, alongside further news data, will be ingested into the pipeline, and become new training data.

Regarding the system architecture, every component is containerised with a separate **Docker** image saved in the **Elastic Container Registry**. The current front-end is implemented using **Streamlit** and we use **flask** for the back-end. In order to optimise the workflow, algorithms developed in **Python** and the **Transformers library** are exposed as a REST service using **Flask**. All model artefacts are in **S3**. In order to orchestrate the Docker containers within the Working nodes of the cluster, we are using **Fargate and ECR**, and we use **MongoDB** as our main database solution. A demo can be found here: https://youtu.be/2btvyEvb8go



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.

Currently, there are two main approaches to **abstractive summarization**:
1) Seq2Seq models (such as T5) and 2) Autoregressive models (such as the GPT family). For the EXPLORE phase, we used the **mT5 and the mbart models** trained on the **Dutch CNN/DailyMail Dataset.** The core of the AI functionality was built on **Python**, and related libs (pandas, NumPy etc), and we used the **transformers library** for the pre-trained models.

Our **conclusion** is that these models can achieve high ROUGE scores, yet we need to adapt the summaries to comply with the journalistic style of VRT. In the EXPERIMENT phase, the objective is to finetune these models on the **VRT dataset,** and if possible using the **factory-ia infrastructure provided by CEA**, since they have the V100 GPUs The objective of fine-tuning will be three-fold:
- **Improve Dutch language generation**: Although most models we tried out during the EXPLORE phase can in fact generate text in dutch out-of-the-box, it is always good to fine-tune on the target language of the application.
- **Optimise for Summarization**: Improve the generalist pre-trained models just on summarisation
- **VRT style transfer**. The training corpus of these models do have journalistic examples, but it is important that we guide the model towards the general style VRT employs, especially as we are focusing on text generation.

The fine-tuned model will be **dockerized** and ready to be deployed by any cloud provider using **Terraform** . As for the API in front of the model, we will work with VRT to define the endpoints on their and needs in terms of protocols(http, grpc) and parameters

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

Technical Scalability: The solution will be provided as **SaaS** (Software as a service) to facilitate replicability. We have created **Terraform** templates, so that it can be deployed in any cloud provider (AWS, Azure, GCP etc), increasing the **flexibility of the solution**. The manifest files define **autoscaling** groups for horizontal scalability depending on the compute/memory/IO load. Similar deployments are used by tech giants serving thousands of API calls per second.

Operational Scalability: Algomo is part of **NVidia's inception program** and we get bespoke consulting specifically on us with scaling AI deployment in the cloud. Our **cloud infrastructure is managed by specialists** (Cloudvisor), who make sure that we follow best practices for cloud scalability, and we have a **dedicated DevOps person** internally. Lastly, our **team of ML engineers** continuously research and optimise the efficiency and scalability of the ML algorithms

Business Scalability: Summarisation is a big domain in NLP, and can be used in multiple use cases, and beyond news articles. One

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

of the reasons we applied in REACH in the first place was because **we use summarisation for our core product**, which is a multilingual customer service automation tool. A key use case for us is scraping the FAQ section of a website and providing short summaries that become the responses of our AI, but also as a paraphrasing engine, that provides additional training data for our AI. The same pipelines we use for VRT can be used for **Media Monitoring, SEO, Social Media Marketing, Legal contract analysis, Financial research** etc.

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 legislation concerning security and privacy (e.g. GDPR).

Data Sources: The data that was used in the pre-trained language models are from **open datasets** which come without any legal restrictions and are available for commercial use. Any subsequent data from VRT are already **public data** that contain no personally identifiable information (PII) data.

Data processing, workflows and standards: Algomo has already implemented the appropriate technical and organizational measures to ensure that our processes meet the **GPDR** and to guarantee the protection of the right of the individuals. We already work with personal information with other customers, and thus everyone within Algomo is committed to confidentiality and Algomo has authorization and control measures in place. Algomo is also accredited by UK **Information Commissioners Office** and fully complies with the Data Protection Act 2018. Our developers have also signed a **confidentiality agreement** that prevents them from sharing any information with third parties.

Technical tools/implementation: all data is encrypted over **HTTPS** on transit and **encrypted at rest** using **KMS**. VRT owns their data and may contact us to extract or delete all their data. Every user-facing service is protected by an **authentication gateway**, where each user can access their own resources using a **personal API key**. At Algomo we use AWS, where access to our **AWS** resources is managed by a **separate IAM account,** and we enforce **2FA authentication.** All our implementations are in a **private Virtual Private Cloud,** where only authorised members and resources can access. From the Reach toolbox, we intend to use the **Anonymizer** to remove any possible PII data, and the **Data Sharing Platform** to exchange data with VRT. The above toolkit ensures a **secure and Trusted data Value Chain**, and we will continue reviewing, monitoring and adapting during the course of the project

5. **QUALITY ASSURANCE AND RISK MANAGEMENT:** Describe the quality process planned 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.

QUALITY PROCESS: For machine summarization, there are two main metrics used to evaluate the performance of a model

Bleu: percentage of n-grams in the machine-generated summaries that appear in the human reference summaries (precision)
Rouge: percentage of n-grams in the human reference summaries that appear in the machine-generated summaries (recall)
In the EXPLORE phase we used the Rouge score. The above two metrics are quite crude, as they ignore synonyms of semantically similar phrases. For that reason in the EXPERIMENT phase, we would like to also use BERTScore, which computes a similarity score for each token in the candidate sentence with each token in the reference sentence, which correlates better with human judgement.

RISKS: (Green stands for Low, Yellow for Medium and Red for High, L: Likelihood, S: Severity, I: Impact)

Risk	L	S	I	Mitigation
DATA SECURITY: Handling sensitive data leading to privacy issues	L	L	L	News data are by their nature public, so there's little (if any) need for anonymization.
DATASET SHIFT: change in the distribution of training data. This could be due to stylistic or domain changes	M	L	L	We will be retraining the model as frequently as possible, but also we will try to use adversarial search to delete any training data that are completely off from the expected distributions
PROJECT MANAGEMENT: Not prioritizing the right tasks, and not putting the right resources for the right tasks	L	H	M	Work closely with VRT to create a product roadmap to clearly scope the work to be done, and what matters the most. We also use Agile internally, so the work done for VRT will be always part of our biweekly sprints.
RESOURCES: Company resources leaving the company or underperforming	M	M	M	Agile practices, and product documentation ensures that knowledge is not lost. We also keep multiple people involved to ensure redundancy.
ML PERFORMANCE: ML models do not get good results.	H	M	H	Use the 80/20 to get the most value as fast as possible. Agile ensures constant delivery.
SCALABILITY: ML models are difficult to train or serve	M	H	H	We use AWS and worst-case scenario we can use Sagemaker. Moreover, we can use the transformers API, which is optimised for these models.

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