Technical Specification Double-side Page

1. **TECHNICAL SCOPE:** Summarize the solution developed during the EXPERIMENT phase: how have you finally addressed the challenge/Theme Challenges and tackled with its requirements and data. Include a diagram.

Task: Developing an abstractive summarization model in Dutch using a dataset provided by VRT

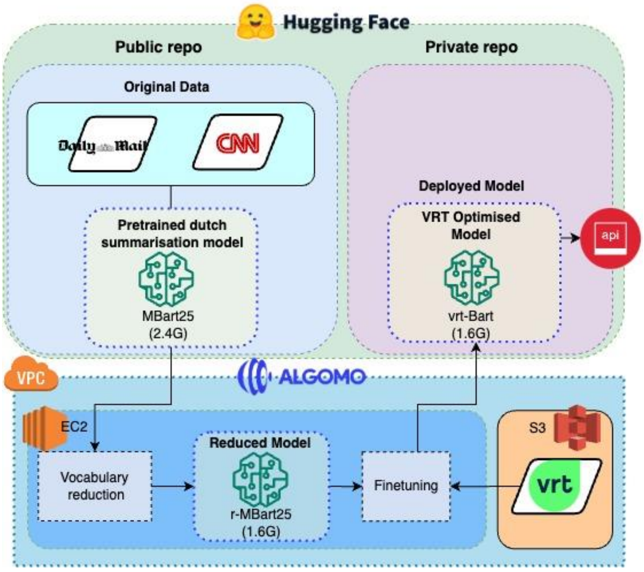
_____________**Process**_____________

1. Base model: We used a pretrained language model for dutch summarisation, freely available from the HuggingFace Library. This is a mBART-25 model, which is multilingual encoder-decoder (sequence-to-sequence) model that had already been finetuned for summarisation in dutch.

2. Model Size Reduction: The above pre-trained model was quite large in size, which made the training process slow and inefficient, as it wouldn't fit into our GPU instances. Moreover, it was not optimised for Dutch. To address this, we **reduced the vocabulary** size to cater only for Dutch tokens.

3. Fine-tuning with VRT data: After reducing the size of the model, we finally finetuned it with **VRT's data**. The final model is deployed and exposed through an API at a private repository in the HuggingFace library, which only VRT can access. Both code and model are in our Gitlab repository.

Full documentation can be found on our notion project page



2. **ALGORITHMS, TOOLS AND CONCLUSIONS:** Detail the algorithms and tools finally selected to accomplish the challenge/Theme Challenges. Summarize the main results that you have obtained during the EXPERIMENT phase: data, insights, conclusions and the main contributions to solve the challenge/Theme Challenges.

Algorithms and Data: We started with a pretrained **mBART-25** model, fine-tuned on the Dutch CNN/DailyMail summarization dataset (290k entries). This model was quite big so according to Abdaoui et al, we decided to reduce the vocabulary size by 80% (from 250k to 52k), which translates to a reduction of 33% of the model parameters. This also reduced the model size from 2.44 GB to 1.6 GB. This has the additional advantage of quicker load and inference times (for example, loading the model has decreased from 7.33s to 3.92s). We then fine-tuned this model with VRT data (150k entries). The training scripts and samples of generated summaries are in a **GitLab repository**, and the model is uploaded to a private **Huggingface** repository.

Tools/Compute: The core of the AI functionality was built on **Python** and related libraries (Pandas, NumPy etc), and we used the **Transformers library** to manage the loading, evaluation and training of the models. We use the Seq2seqTrainer for training. We use **Weights&Biases** to monitor the performance of all models during training and compare the performance of different models in different datasets. We use **AWS** for compute and storage, and all data was in a **private VPC**.

Evaluation/Results: VRT provided us with a private mBART model that they trained on their data. We evaluate the models using the standard metrics in the literature for evaluating abstractive summarization: **Rouge, BertScore and Meteor**. In the test VRT dataset, our new model **outperforms their mode**l as well as three other publicly available models by more than **5 Rouge-1 points**. We outperform the VRT baseline in two additional datasets (XLSUM,CNN) of Dutch summarization. Our model also reduces by **half inference time** after decreasing the vocabulary size. As our experiments verify, the technique we use to perform this reduction also does not affect the performance of the model on Dutch data. More crucially, other ways of reducing the size of a model, such as knowledge distillation, cannot assure that performance will be the same

Main achievements: A **lightweight** (60% smaller), **faster** (2x) and **more performant** model (5 Rouge-1 points).

This showcases that it is possible to combine (1) the generative ability in Dutch of mBART-25 (2) good summarization ability, obtained during pre-training on the Dutch CNN/DailyMail pre-training (3) proper use of VRT style.

3. **SCALABILITY AND FLEXIBILITY OF THE SOLUTION:** Explain how the solution copes with the challenge/Theme Challenges requirements and how can it be adapted to other similar problems. What work is still pending to create a real/stable product if any? What TRL level is it in?

Technical scalability and fit for VRT: VRT expects to be generating around 100-200 summaries daily. Our model, once deployed in a GPU can generate **100 summaries in 66 seconds**, so we can deal with way larger loads. If necessary, the solution can easily scale horizontally, depending on the compute/memory/IO load. Similar deployments are used by tech giants serving thousands of API calls per second. Since the model is in a Huggingface repository, it is easy to integrate with any existing service that uses Python though the Transformers library API. Its reduced size makes it capable to run on small cloud instances and the reduced inference time means it can be used in near real-time applications.

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 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.

TRL/ Next steps: We consider this work to be of **TRL-6**. We need to **test it in operational environments** before fully scaling it to more use cases and clients, and also build model drifting and feedback pipelines that would ensure the performance of the summarisation models won't degrade.

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

Data Sources: The data that was used in the pre-trained language models proceeds from **open datasets** which come without any legal restrictions and are available for commercial use. The private dataset we used from VRT contains no personally identifiable information (PII) data and has been stored in AWS S3 buckets.

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.

Sharing our model: The code has been shared with VRT through a Gitlab private repository. The model has been uploaded as a private model to Huggingface.

5. **QUALITY ASSURANCE AND RISK MANAGEMENT:** Describe the quality process followed for the final product. Technologically, which problems have you encountered and how you have solved them, and any processes followed that guarantee that the solution fulfills the challenge/Theme Challenges and data provider requirements.

QUALITY PROCESS: All models have been evaluated in three datasets: (a) XLM (b) CNN (publicly available Dutch summarization datasets) and (c) the one provided by VRT. On all of these datasets, we evaluate the (a)VRT baseline (b) two publicly available models and c) our model. We also have provided VRT with generated summaries, so that they can perform a human evaluation. The metrics we use have been widely validated in the summarization literature: **Rouge, BertScore and Meteor**.

The results of the evaluation suggest that the model is **better than the VRT baseline**, and it is on-par in datasets it has not been finetuned in with models that were finetuned in those datasets. Through our extensive evaluation process, we are confident our model is better than existing solutions for Dutch summarization, both in terms of the quality of the summaries generated and the improvement in deployability, as the model is smaller and faster.

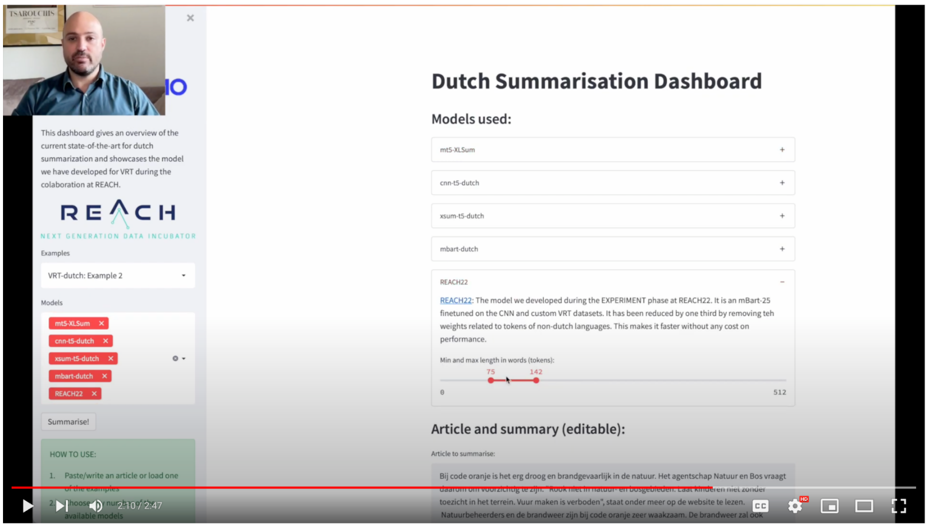
RISKS: (Green stands for Low, Yellow for Medium and Red for High, L: Likelihood, S: Severity, I: Impact)

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 have fine tuned the model following the VRT style of summaries, which has been consistent across extended periods of time. We have eased fine tuning with the reduced model so that it's faster to adapt the model to new topics or styles.
POOR GENERALISATION: evaluation does not correlate with real-world performance	H	M	H	We have used three different metrics during evaluation and three different datasets. Results have been consistent through all the different combinations.
SCALABILITY: ML models are difficult to train or serve	M	H	H	We reduce the model to reduce training and inference times. Use the transformers API, optimised for the model.

Annex 1. Means for accessing the MVP

Please, indicate in 1 page indicating the means for accessing the MVP for a potential customer (login information, website address, link to a demo video or whatever means are needed to check that the MVP exists and works).

→Youtube video showcasing technology and demo←


This dashboard gives an overview of the current state-of-the-art for dutch summarization and showcases the model we have developed for VRT during the colaboration at REACH.

HOW TO USE:

1. Paste/write an article or load one of the examples
2. Choose any number of the available models

Dutch Summarisation Dashboard

Models used:

- mt5-XLSum +
- cnn-t5-dutch +
- xsum-t5-dutch +
- mbart-dutch +
- REACH22 −

REACH22: The model we developed during the EXPERIMENT phase at REACH22. It is an mBart-25 finetuned on the CNN and custom VRT datasets. It has been reduced by one third by removing teh weights related to tokens of non-dutch languages. This makes it faster without any cost on performance.

Min and max length in words (tokens): 75 142 (0 – 512)

Article and summary (editable):

Article to summarise:

Bij code oranje is het erg droog en brandgevaarlijk in de natuur. Het agentschap Natuur en Bos vraagt daarom om voorzichtig te zijn. "Rook niet in natuur- en bosgebieden. Laat kinderen niet zonder toezicht in het terrein. Vuur maken is verboden", staat onder meer op de website te lezen. Natuurbeheerders en de brandweer zijn bij code oranje zeer waakzaam. De brandweer zal ook

https://www.youtube.com/watch?v=uEViGkBYE1g