Extractive & Abstractive Auto Summarization (NL)

Amberscript

vrt REACH

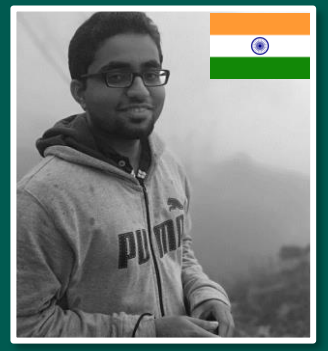
Meet the team

Peter-Paul de Leeuw
Business Lead

Peter-Paul holds a MBA from INSEAD and cum laude master's in Economics from the Erasmus University of Rotterdam. He has several years' experience working in strategy consultancy. Within Amberscript Peter-Paul is responsible for Strategy and product development.

Timo Behrens
Technical Lead

Timo holds a degree in Electrical Engineering from TH Köln. As technical lead Timo is responsible for overall development, making sure Amberscript deploys state of the art technology with optimal reliability and quality.



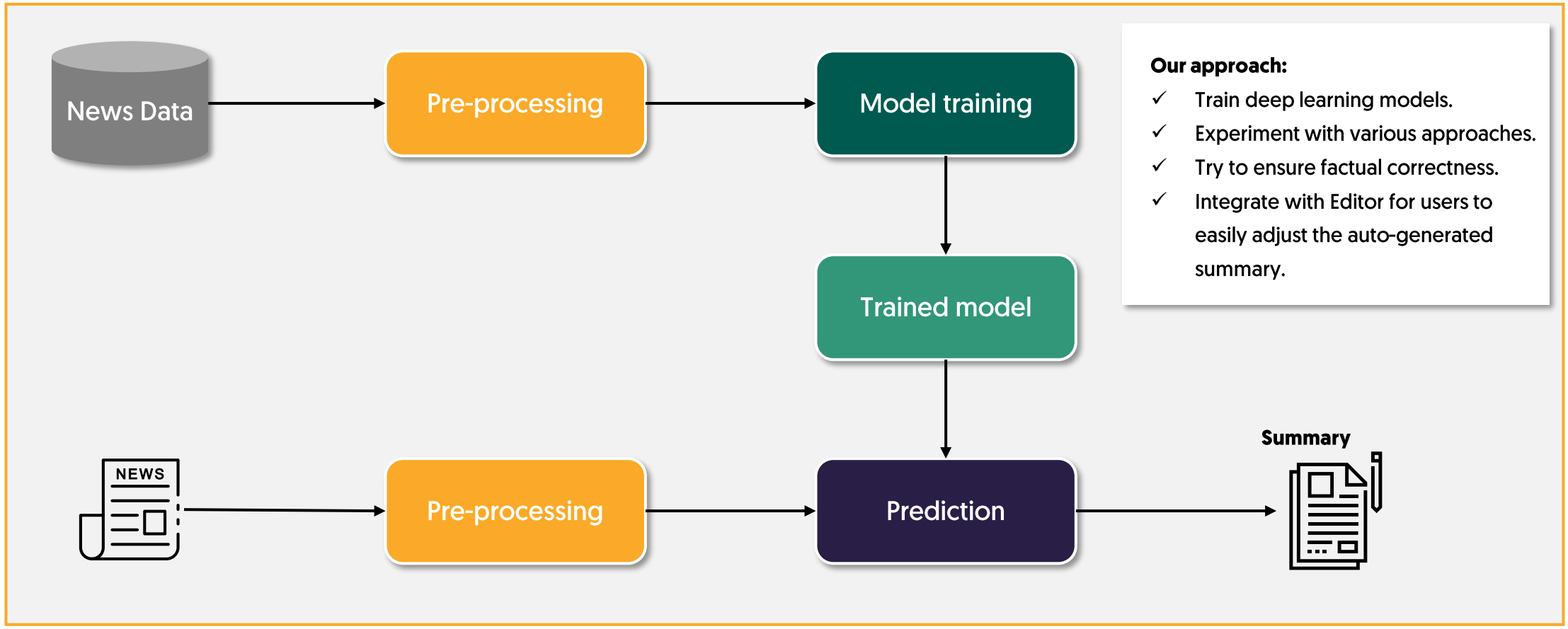
Nithin Holla
ML Engineer

Nithin holds a master's degree (cum laude) in Artificial Intelligence from the University of Amsterdam. He is responsible for research and development of Amberscript's natural language processing and speech recognition components.

Jolien de Louw
Project Manager

Jolien holds a master in Economics from Utrecht University and has 8 years of experience as a management consultant at Accenture. She is responsible for operations management and customer support at Amberscript.

Our technical proposal in a nutshell



Our approach:

- ✓ Train deep learning models.
- ✓ Experiment with various approaches.
- ✓ Try to ensure factual correctness.
- ✓ Integrate with Editor for users to easily adjust the auto-generated summary.

Two types of summarization models

1 EXTRACTIVE SUMMARIZATION

Extractive summarization involves concatenating important sentences taken from the article into a summary.

Specifications:

- VRT data consists of abstractive summaries.
- Unsupervised extractive summarization

2 ABSTRACTIVE SUMMARIZATION

Abstractive summarization involves generating novel sentences from information extracted from the article.

Specifications:

- VRT data consists of abstractive summaries.
- Supervised abstractive summarization.

1. Deep dive into extractive summarization

1 EXTRACTIVE SUMMARIZATION

Our step-by-step approach:

1. Create a graph with sentences as nodes.
2. Initialize node representations (experiment with 3 approaches):
    a. TF-IDF.
    b. Representations from pretrained BERTje.
    c. Pretrain BERTje on VRT data.
3. Set edge weights using a similarity metric.
4. Run PageRank to obtain sentence importance scores.
5. Sort the sentences according to scores.
6. Trigram blocking to increase diversity.
7. Create extractive summary.
8. Evaluate output with ROUGE metric and language experts.

Implementation:

- Python
- PyTorch
- Hugging Face transformers

2. Deep dive into abstractive summarization

2 ABSTRACTIVE SUMMARIZATION

Context:

- T5 - text-to-text transformer
- mT5 - multilingual T5
- Pretrained on 101 languages

Step-by-step approach:

1. Fine-tune mT5 to generate summaries:
    a. VRT news articles as inputs.
    b. Human-generated summaries as targets.
2. Mitigate factual inconsistencies with pointer generator networks.
3. Create abstractive summary.
4. Evaluate output with ROUGE metric and language experts.

Implementation:

- Python
- PyTorch
- Hugging Face transformers

Two-stage summarization

We will also experiment with combining the extractive and abstractive summarization models:

1. Select important sentences with unsupervised extractive summarization.
2. Generate abstractive summary.
3. Evaluate output with ROUGE metric and language experts

This approach is helpful for long documents (>1000 words).

Model evaluation & quality assurance

PROJECT MEASURES

- In case of Dutch data scarcity, translate English summarization datasets to Dutch.
- Factual inconsistency will be mitigated by employing pointer generator networks, which allows copying of information from the source article along with generating novel sentences.

QUALITY ASSURANCE

- Evaluation using ROUGE metric.
- Our 500+ native freelance network of language experts.

DATA SECURITY & SERVER UTILIZATION

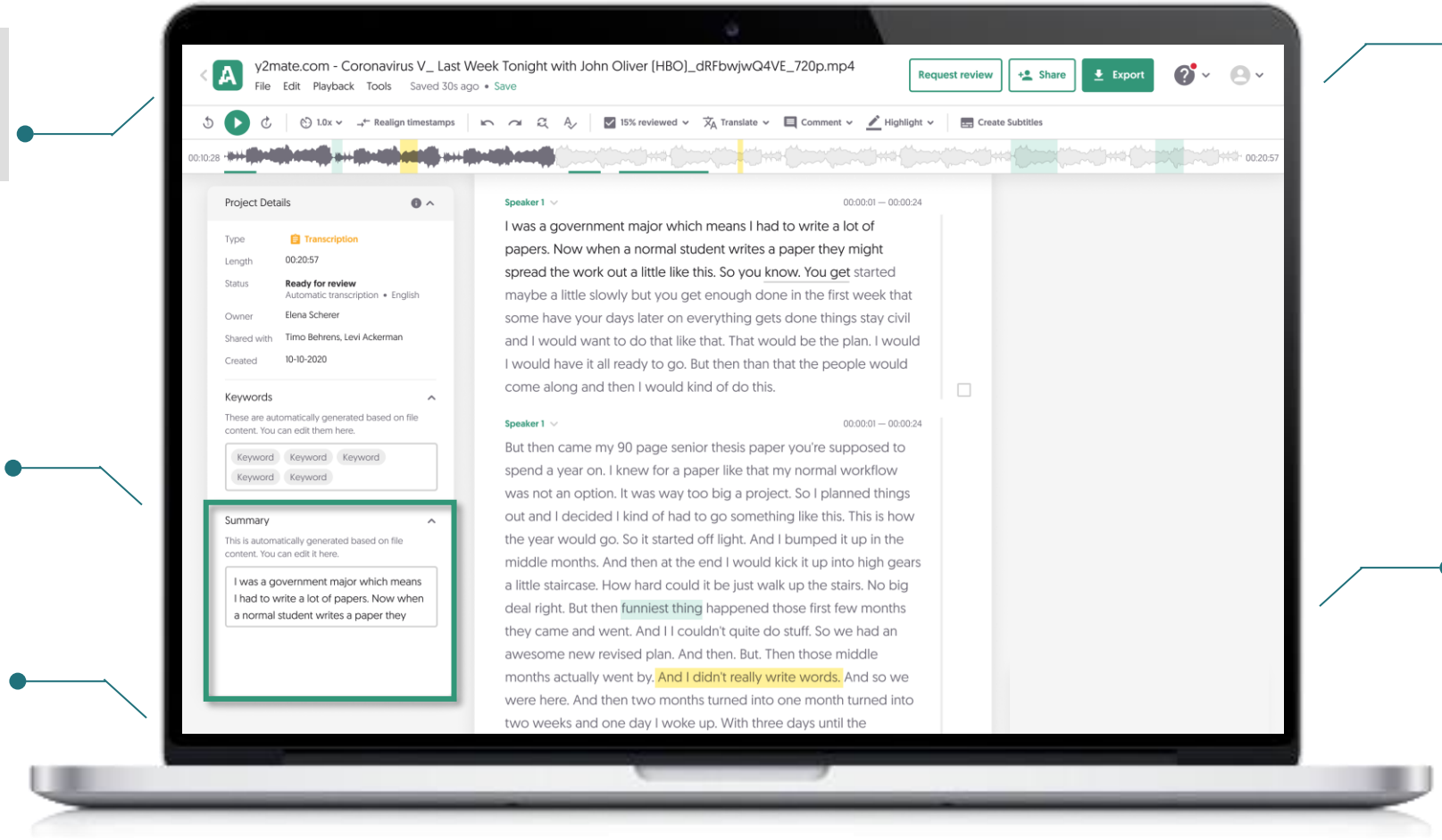
- For training the models, we only use news data that is already publicly available. Thus, it includes no personal data and does not violate the privacy of any individual/ organization. Next to that, we adhere to the GDPR *Privacy by Design* principle.
- Data received from VRT will be treated confidentially.
- All project members adhere to strict privacy and security standards and undergo regular training in IT security and privacy.
- Data is stored with a leading cloud-provider on servers in Europe. The servers are certified according to ISO27001 and ISO9001 standards.

An integrated editor to modify generated summaries

Export functionality
Quickly export the summary in Text, Docx and many other formats.

Summary
The automatic generated summary based on the file content is shown. It can easily be edited by the user.

Auto Save
All changes made are directly and automatically saved



Personal login
Users can login using 2-factor authentication.

Highlight
Highlight proper names, numbers so that journalists can check for consistency.

QUESTIONS

Amberscript B.V.
Keizersgracht 668B
1017 ET Amsterdam
the Netherlands

Commercial Register (KvK): 70427585
VAT (BTW) nr: NL858314071B01

Telephone: +31 20 24 40 992
Availability: Mon – Fri 09:00 – 18:00
Contact by email: info@amberscript.com