Mauro Giovanoli
mauro.giovanoli@srf.ch

Automating Video Metadata Generation: Evaluating RNN, BERT, and GPT-4 for Efficient Tagging

Research Objective:

This CAS thesis investigated the potential for **automatic tagging of video content at SRF** by exploring three state-of-the-art approaches: a traditional Recurrent Neural Network (**RNN**), a fine-tuned Bidirectional Encoder Representations from Transformers (**BERT**) classifier, and OpenAI's Generative Pre-trained Transformer (**GPT-4**).

Data:

- Every video item in Play SRF can be accessed and viewed separately on Play SRF.
- In the example above, an item corresponds to one of several "Tagesschau" stories.
- Each item has its own descriptive **web-title** and **web-lead** which describes the content briefly.
- These manually curated descriptions need to be as short as possible while still describing as accurately as possible what our users can expect from the video.

Stream Now

...
5. Video Item
   web-title: ...
   web-lead: ...
6. Video Item
   web-title: Swiss Steel: Aktionäre stimmen für ...
   web-lead: Swiss Steel sitzt auf einem Schuldenberg von fast einer Milliarde Euro. An der ausserordentlichen Generalversammlung haben die Aktionäre nun ...
7. Video Item
   web-title: Berlin: Fliegerbombe aus dem zweiten Weltkrieg...
   web-lead: Spezialisten haben in Berlin eine rund 100 Kilogramm schwere Fliegerbombe aus dem zweiten Weltkrieg unschädlich gemacht...
8. Video Item
   web-title: ...
   web-lead: ...

BERT ?

RNN ?

GPT-4 ?

SRF Archive

- At SRF Archive the video items are **manually** tagged with up to two descriptors from a controlled vocabulary (n=26) to enhance archive discoverability.
- These descriptors describe the **main topics** of an item, focusing on the **abstract information** rather than the visual aspects.

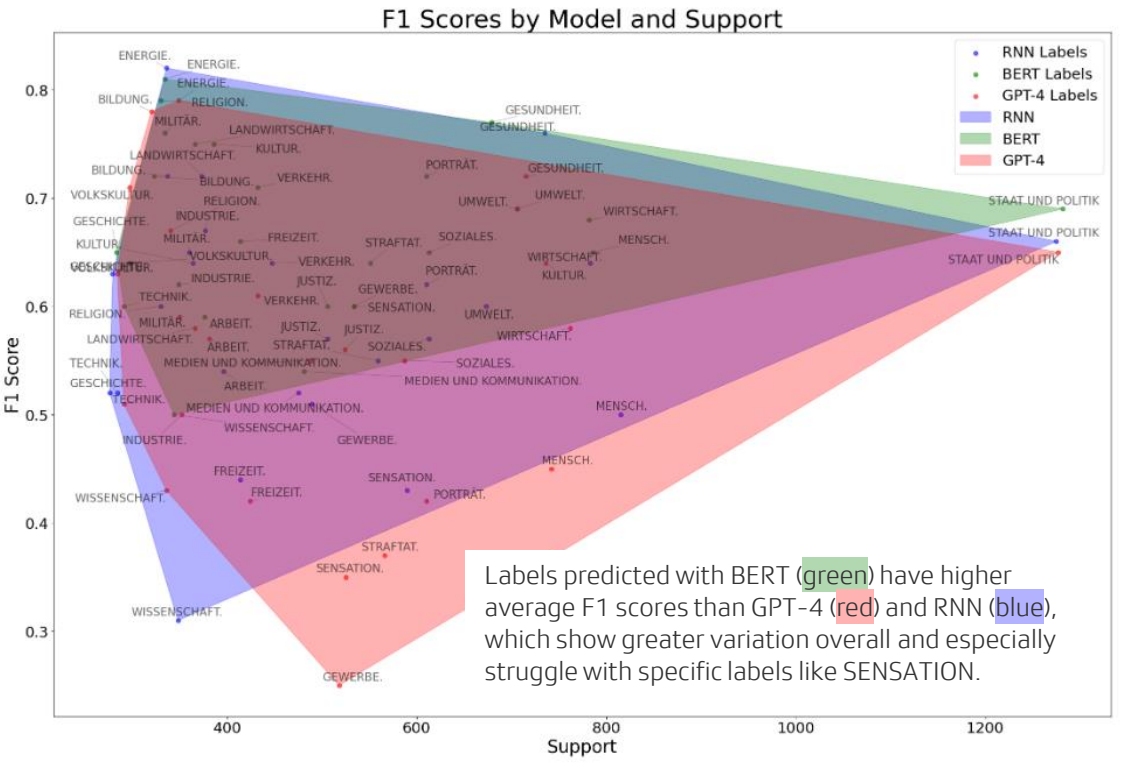
Results:

And the winner is: **BERT**

	Precision	Recall	F1-Score	Accuracy
1. BERT	0.74	0.59	0.65	95.1%
2. RNN	0.75	0.47	0.57	94.7%
3. GPT 4	0.65	0.5	0.54	93.7%

While all three models perform similarly to human tagging in predicting top-level descriptors, the fine-tuned BERT classifier, though comparable in precision, outperforms the other two models in recall.

This means that, in practical terms, the BERT classifier is more effective (high accuracy) at identifying all relevant descriptors, ensuring fewer relevant items are overlooked in the automated tagging process.



Labels predicted with BERT (green) have higher average F1 scores than GPT-4 (red) and RNN (blue), which show greater variation overall and especially struggle with specific labels like SENSATION.

Impact:

Remarkably, in addition to accurately categorizing videos by short description and title, the final BERT model has also stood out in categorizing SRF articles by lead and title, demonstrating **outstanding results** and **substantial potential towards its practical application**. BERTs exceptional generalization capability makes it highly adaptable to various metadata tasks and positions it as an ideal solution for increasing SRF's efficiency and accuracy in generating content metadata through automation.