Fine-tune a Small Language Model (SLM) for Summarization

AUTHORS

Zeinab GHAMLOUCH
Alessa MAYER

2025-03-14

1 Introduction

In this project, we focus on Fine-tuning a Small Language Model for the task of text summarization in a non-English language. While large language models (LLMs) have shown impressive performance in summarization tasks, there is growing interest in developing smaller, more efficient models that can achieve comparable results with fewer computational resources. In the following, we explain every step taken to achieve this goal, from dataset preparation and synthetic summary generation (2) to model fine-tuning (3) and evaluation (4). We also highlight the challenges of working with non-English text and demonstrate how SLMs can be effectively applied to summarization tasks in resource-constrained environments. We finish with a discussion of the results 5 and, finally, we conclude our findings 6.

2 Dataset

2.1 Dataset Description

For this project, we used the "German Political Speeches Corpus" [1], a publicly available dataset sourced from Kaggle. The dataset consists of 6,685 speeches delivered by 71 speakers, spanning the period from 1984 to 2017 (mostly from 1990 onwards). These speeches were primarily held by top German officials, including presidents, chancellors, and members of the Bundestag, and were selected based on their political relevance. The texts were collected from official government sources, such as the German Presidency, Chancellery, Bundestag, and Ministry of Foreign Affairs, as well as personal archives of prominent figures like Helmut Kohl, Wolfgang Thierse, and Norbert Lammert. The dataset is provided in XML format, with each file containing both the speech text and associated metadata.

We combined the data sources into a single dataframe and focused solely on the "rohtext" column, which contains the raw text of the speeches. The rest of the metadata was removed to create an unannotated dataset suitable for our analysis. Additionally, we removed the last 1,685 speeches to reduce the dataset to a total of 5,000 speeches, as required by the project description.

2.2 Dataset Annotation

To generate high-quality synthetic summaries for the German political speeches, we used a large language model (LLM), specifically Llama 3.2 [2], as the baseline model. Llama 3.2 was chosen for its ability to produce coherent and contextually accurate summaries, even for complex and domain-specific text like political speeches. Moreover, it can achieve good performance even when dealing with a foreign language like German. The summaries generated by this LLM serve as the gold standard for training and evaluating our small language model (SLM).

The main challenge we had to tackle is finding the perfect prompt to get the best summarization possible to ensure that our SLM will be trained and evaluated on a stable baseline. Thus, we tried different prompts and manually evaluated the first five results. We found that having the whole prompt in German, instead of giving instructions in English, yielded the best results. The final prompt we used can be found here 7.1.

2.3 Dataset Splitting

We used the train_test_split function from the Scikit-Learn library (Pedregosa et al., 2011) [3] to split our dataset into training, test and validation sets and decided to use the following distribution of the data:

- 70 % for training
- 15 % for testing
- 15 % for validation

3 Methodology

3.1 Model Selection

As our Small Language Model we chose the DistilBART model and more specifically "sshleifer/distilbart-cnn-12-6" [4]. Its smaller size ensures that it can run efficiently on Colab's free-tier GPU, and it meets the criteria of a Small Language Model (<7B). Despite its compact design, DistilBART retains much of the performance of the original BART model, especially for summarization tasks. It has been fine-tuned on the CNN/DailyMail dataset, which equips it with strong capabilities for generating concise and coherent summaries. For our specific task of summarizing German political speeches, DistilBART's ability to process long-form text and extract key information aligns well with the nature of political discourse, which often involves complex sentences and nuanced arguments. Though pre-trained on English text, DistilBART's architecture is still well-suited for adaptation to German. In the following, we will explain how we finetuned the chosen model further.

3.2 Model Fine-Tuning

The fine-tuning process for our German political speeches summarization task was implemented using the Hugging Face Transformers library. We utilized the "Trainer" class to streamline the training process. The "TrainingArguments" class was used to define the training configuration, including key hyperparameters such as learning rate, batch size, and the number of training epochs. Additionally, we employed a "DataCollatorForSeq2Seq" to handle padding and ensure that input sequences were properly aligned for the encoder-decoder architecture of DistilBART.

Moreover, we utilized the Bitsandbytes library for quantization and LoRA (Low-Rank Adaptation) [5] for parameter-efficient fine-tuning. These techniques allowed us to adapt the DistilBART model effectively while minimizing memory usage and computational costs. In our setup, we configured Bitsandbytes to store the model in 4-bit precision using the nf4 (Normalized Float 4) quantization method. We also enabled double quantization to further reduce memory overhead. The computations were performed using bfloat16 (BF16), a 16-bit floating-point format that balances precision and efficiency. To further optimize the fine-tuning process, we applied LoRA and, more specifically, the PEFT (Parameter-Efficient Fine-Tuning) library to implement LoRA. We found the following parameters in the LoRA configuration to work best: r=8, "lora_alpha"=16 (2 * rank), "lora_dropout"=0.05, "target_modules"=["q_proj", "v_proj", "k_proj", "out_proj"], "bias"="none", "task_type"="SEQ_2_SEQ_LM".

During fine-tuning, we experimented with several hyperparameters to optimize the model's performance. The learning rate was set to $2e - 5$, which is a common choice for fine-tuning pre-trained models, as it balances the need for adapting to the new task while avoiding catastrophic forgetting of the pre-trained knowledge. Using smaller learning rate decreased the loss during training. We used a batch size of 8 for both training and evaluation, which was the maximum feasible size given the memory constraints of Google Colab's free-tier GPU. The model was trained for 6 epochs, which provided sufficient time for the model to learn task-specific patterns without overfitting. Using only 3 epochs gave less performance, so we increased it to 6. Weight decay of 0.01 was applied to regularize the model and prevent overfitting.

In summary, the fine-tuning process leveraged the Hugging Face Transformers library to efficiently adapt DistilBART to our summarization task. By carefully selecting and tuning hyperparameters, we were able to achieve a balance between computational efficiency and model performance, ensuring that the fine-tuned model could generate summaries of German political speeches compatible with our baseline model.

3.3 Evaluation Metrics

To evaluate our fine-tuned DistilBART model for German political speech summarization, we employed a multi-faceted approach using ROUGE [6], BERTScore [7], and LLM-as-a-Judge. These metrics collectively provide a comprehensive assessment of summary quality, balancing lexical, semantic, and human-like evaluation.

- **ROUGE** measures n-gram overlap between generated and reference summaries. We used ROUGE-1 (unigrams), ROUGE-2 (bigrams), and ROUGE-L (longest common subsequence) to assess lexical similarity and coherence.

- **BERTScore** evaluates semantic similarity using contextual embeddings from the 'bert-base-multilingual-uncased' model. It computes precision, recall, and F1 scores, offering insights into how well the generated summaries align with the reference summaries in meaning.

- **LLM-as-a-Judge** provides a human-like evaluation by comparing generated summaries to reference summaries. We used Llama 3.2 to score each generated summary on four criteria - relevance, coherence, conciseness, and fluency - on a scale of 1 to 5. The scores were averaged across the test dataset and normalized to a range of 0 to 1 for each category.

These metrics complement each other: ROUGE focuses on lexical overlap, BERTScore captures semantic alignment, and LLM-as-a-Judge evaluates higher-level aspects like coherence and fluency. Together, they ensure a robust evaluation of summary quality, confirming that our model generates concise, accurate, and contextually appropriate summaries of German political speeches. We computed average scores across the entire test dataset for all metrics, providing a comprehensive and aggregated measure of the model's performance.

4 Results

The fine-tuned model's performance on the test set was evaluated using several metrics, including ROUGE scores, BERTScore, and LLM-based evaluation criteria. The scores can be found in Table 1

Table 1: Evaluation Metrics for the Fine-Tuned Model

Metric	**Score**
ROUGE Scores	
ROUGE-1 F1	0.3335
ROUGE-2 F1	0.0828
ROUGE-L F1	0.1721
BERTScore	
BERT Precision	0.6764
BERT Recall	0.6748
BERT F1 Score	0.6752
LLM Evaluation Metrics	
Relevance	0.455
Coherence	0.5392
Conciseness	0.465
Fluency	0.4343

The ROUGE-1 F1 score of 0.3335 indicates moderate overlap between the generated summaries and the reference texts at the unigram level. The ROUGE-2 F1 score of 0.0828 is relatively low, suggesting limited overlap at the bigram level, which implies that the model struggles to capture sequential word pairs effectively. The ROUGE-L F1 score of 0.172 reflects modest performance in matching the longest common subsequences, indicating room for improvement in generating coherent and contextually aligned summaries.

The BERTScore metrics show strong performance, with a precision of 0.6764, recall of 0.6748, and an F1 score of 0.6752. These results suggest that the model achieves high semantic similarity between the generated and reference texts, even if the surface-level overlap (as measured by ROUGE) is lower.

Last but not least we are going to analyze the LLM Evaluation Metrics. The relevance score of 0.455 indicates that the generated text is moderately relevant to the context. With a score of 0.5392, the model performs reasonably well in maintaining consistency in the generated text. Considering conciseness, the score of 0.465 suggests that the model could improve in producing more concise summaries without unnecessary details. The fluency score of 0.4343 indicates that the generated text is somewhat fluent but may still contain grammatical or stylistic issues.

In conclusion, the model demonstrates strong semantic understanding, as evidenced by the high BERTScore, but struggles with surface-level metrics like ROUGE-2 and ROUGE-L. While coherence is relatively strong, improvements in relevance, conciseness, and fluency could enhance the overall quality of the generated text. The low ROUGE-2 score highlights a specific area for improvement: capturing sequential relationships between words more effectively.

5 Discussion

5.1 Analyses

The evaluation results highlight both strengths and weaknesses of the fine-tuned model. While the BERTScore F1 metrics demonstrate strong semantic alignment, indicating the model effectively captures the essence of the text, the low ROUGE-2 score reveals a significant limitation in capturing sequential relationships between words. This suggests the model struggles with maintaining logical flow and coherence in summaries. The LLM-based evaluation further underscores areas for improvement, particularly in relevance, conciseness, and fluency, indicating that the summaries could better focus on key points, avoid unnecessary details, and improve grammatical quality.

5.2 Limitations

Our approach faced several limitations. The dataset size, while adequate, could be expanded to include more diverse political speeches to improve generalization. The model's size (DistilBART) constrained its ability to handle complex, long-form text effectively. Hardware limitations, particularly the memory constraints of free-tier Google Colab, restricted our ability to experiment with larger models like mBART or more advanced techniques like reinforcement learning. Finally, the lack of human feedback during evaluation limited our ability to refine the model further based on real-world usability.

5.3 Future work

To improve the model further, several steps can be taken:

- Experiment with larger models: Transitioning to larger models like mBART or T5 could enhance summarization quality by leveraging their greater capacity for understanding complex text.

- Post-processing techniques: Implementing reinforcement learning or post-training methods could improve summary coherence and detail, addressing some of the shortcomings identified by the LLM-as-a-Judge evaluation.
- Larger and more diverse datasets: Expanding the dataset to include a wider variety of political speeches and languages could improve the model's generalization capabilities.
- Human feedback integration: Incorporating human feedback into the evaluation and fine-tuning process would provide more nuanced insights and help refine the model for real-world applications.
- Advanced techniques: Exploring multi-turn learning or domain adaptation techniques could further enhance the model's ability to handle the complexities of political discourse.

By addressing these areas, future work can build on the current model's strengths while overcoming its limitations, ultimately leading to more accurate and contextually appropriate summarization of German political speeches.

6 Conclusion

The fine-tuned DistilBART model demonstrated strong semantic understanding, as evidenced by high BERTScore metrics, but struggled with surface-level lexical overlap, particularly in capturing sequential word relationships (low ROUGE-2 score). While coherence was relatively strong, improvements in relevance, conciseness, and fluency are needed. The model did not outperform the baseline Llama 3.2 in all aspects, but it showed promise for efficient summarization in resource-constrained environments.

Fine-tuning SLMs for summarization is crucial, especially for non-English text, as it offers a balance between performance and computational efficiency. This work highlights the potential of SLMs in practical applications, such as summarizing German political speeches or news articles for quick reading, particularly in settings where computational resources are limited. Future work could focus on larger models, diverse datasets, and advanced techniques to further enhance performance.

References

[1] Mexwell, "German political speeches corpus." `https://www.kaggle.com/datasets/mexwell/german-political-speeches-corpus`. Retrieved March 3, 2025.

[2] A. Meta, "Llama 3.2: Revolutionizing edge ai and vision with open, customizable models." Meta AI Blog, September 2024. Retrieved March 10, 2025.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, and G. Louppe, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, 01 2012.

[4] S. Shleifer, "sshleifer/distilbart-cnn-12-6." `https://huggingface.co/sshleifer/distilbart-cnn-12-6`. Retrieved March 10, 2025.

[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *CoRR*, vol. abs/2106.09685, 2021.

[6] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004.

[7] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," 2020.

7 Appendix

7.1 Prompt used to generate summaries

The prompt we used to generate our baseline summary:

```
Fasse den folgenden Text zusammen: {text}
```

This can be loosely translated to "Summarize the following text: {text}". Its simplicity ensured a good baseline.

7.2 Prompt used to evaluate performance (LLM-as-a-Judge)

The prompt we used to evaluate the generated summaries and enable comparison to the baseline:

```
You are an expert evaluator for summarization tasks. Compare the generated summary to
the reference summary based on the following criteria:
    1. Relevance
    2. Coherence
    3. Conciseness
    4. Fluency

    Generated Summary: {generated_summary}
    Reference Summary: {true_summary}

    Provide a score for each criterion on a scale of 1 to 5 (1 = poor, 5 = excellent).
    Return the scores as follows
    [relevance_score, coherence_score, conciseness_score, fluency_score].
    Do not return anything else but this list with scores.
    Example: [0.5, 5, 4.5, 3].
```

We tried to also use a German prompt for evaluation, but found that the returned numbers were not always meeting the given format, while with the prompt above we only retrieved correct sized lists.

7.3 Summary example

Original Text: Guten Tag, Bautzen! Danke fuer die freundliche Begruessung. Mit Blick auf den Ministerpraesidenten natuerlich auch: Guten Tag, Sachsen! Sie wissen ja, dass ich Mecklenburger bin, und wenn ich Sachsen so freundlich begruesse, dann hat das historische Gruende. Auf die komme ich gleich zu sprechen. Aber zunaechst will ich mich bedanken, dass Sie alle meiner Einladung gefolgt sind. Danken moechte ich auch der Hausherrin. Hier im sorbischen [...]

Generated Summary: Dieser Text beschreibt eine Begruessung des Ministerpraesidenten in der Deutschland und Bautzen. Er betont die Bedeutung von einer Einladung von der Hausherrin, die sich auch auch die Herausforderungen zu werden.

Der Text betont, dass sich sich eine unter einer Ueberschrift versammeln koennen, das unterschiedliche Milieus mischen, und sich untersichtliche Ansichten vertreten sind. Der Text ist ein wichtig auch fuer die Eigenstaendigkeit dieser Region stehen. Er erwaehnt auch ein Buergermeister mit Migrationshintergrund, der kommt naemlich aus Berlin, und das ist in dieser Gegend jaht nicht selbstverstuendlich, dann hier Buergershinterung zwischen Berlin zu sein. Der Text betont die Begehaltung von die Husherrin und dass die Hausfahrer, die ihr Nachbar, die schon seit Jahrzehnten hier lebt, und ihr ist hierhergezogen ist. Insgesamt dankt der Text einer Aufstuetzung von den Hausferrin und die Haufern, die dass ihre Einluendern wird. Er bedankt sich dass ein Einladenung gefolgt sind, sondern auf die Haendern und ihrer Wirtschaft zu beziehungen. 3. Die Hausverfahrung der Haufen zu entwickeit, um dass das Zusammenarbeit zu staerken. 4. Die Text erhaet die Huenderung von das Hauserin, die Baufuehlt, die Gewerkschafterin, der Fussballtrainer, die Abiturientin und Berufsschueler, die Frau, die

Baseline Summary: Dieser Text ist ein politischer Redeschritt des saechsischen Ministerpraesidenten Michael Kretzelsberger, in dem er die Bedeutung von Dialog und Offenheit im demokratischen Diskurs betont. Er spricht ueber die Herausforderungen, denen Deutschland und Sachsen gegenueberstehen, insbesondere im Hinblick auf Fluechtlingspolitik und europaeische Integration.

Kretzelsberger fordert dazu auf, seine Meinung einzubringen und sich einzumischen, um gemeinsam an der Gestaltung des Gemeinwesens zu arbeiten. Er moechte, dass Menschen nicht nur Zuschauer sind, sondern aktiv teilnehmen koennen. Er betont die Wichtigkeit von Konfliktloesung durch offene Diskussion und sucht den "offenen Wort" fuer eine solche Debatte.

Der Text ist insgesamt ein Aufruf zur politischen Beteiligung und zum engagierten Engagement in der Demokratie. Kretzelsberger moechte, dass Menschen ihre Meinungen ausdruecken koennen und sich mit anderen austauschen, um gemeinsam Loesungen fuer die Herausforderungen des 21. Jahrhunderts zu finden.