Prompt-Based Approach for Czech Sentiment Analysis

Jakub Šmíd[*] and **Pavel Přibáň**[*,†]

University of West Bohemia, Faculty of Applied Sciences, Czech Republic
[*]Department of Computer Science and Engineering,
[†]NTIS – New Technologies for the Information Society,
{jaksmid,pribanp}@kiv.zcu.cz
http://nlp.kiv.zcu.cz

Abstract

This paper introduces the first prompt-based methods for aspect-based sentiment analysis and sentiment classification in Czech. We employ the sequence-to-sequence models to solve the aspect-based tasks simultaneously and demonstrate the superiority of our prompt-based approach over traditional fine-tuning. In addition, we conduct zero-shot and few-shot learning experiments for sentiment classification and show that prompting yields significantly better results with limited training examples compared to traditional fine-tuning. We also demonstrate that pre-training on data from the target domain can lead to significant improvements in a zero-shot scenario.

1 Introduction

In recent years, pre-trained BERT-like (Devlin et al., 2019) models based on the Transformer (Vaswani et al., 2017) architecture and language modelling significantly improved the performance of various NLP tasks (Raffel et al., 2020). The initial approach was to pre-train these models on a large amount of text and then fine-tune them for a specific task. More recently, an approach exploiting the nature of language modelling appeared, called *prompting* or *prompt-based fine-tuning*. Prompting is a technique that encourages a pre-trained model to make specific predictions by providing a prompt specifying the task to be done (Liu et al., 2023).

This new approach became very popular in solving NLP problems in zero-shot or few-shot scenarios, including sentiment analysis (Gao et al., 2021, 2022; Hosseini-Asl et al., 2022). Most of the current research aimed at languages other than Czech, especially English. To the best of our knowledge, no research has focused on any sentiment analysis task in the Czech language using prompt-based fine-tuning. To address this lack of research, this paper presents an initial study focusing on two sentiment-related tasks: **aspect-based sentiment analysis** and **sentiment classification** in the Czech language by applying prompt-based fine-tuning.

The *sentiment classification* (SC), also known as *polarity detection*, is a classification task where the objective for a given text is to assign one overall sentiment polarity label. Usually, the three-class scheme with *positive*, *negative* and *neutral* labels is used, but more labels can be applied (Liu, 2012).

Aspect-based sentiment analysis (ABSA) is a more detailed task compared to SC, which aims to extract fine-grained information about entities, their aspects and opinions expressed towards them. Generally, the goal of ABSA is to identify the sentiment of each aspect or feature of a product or service. There are multiple definitions and versions of the ABSA task (Pontiki et al., 2014; Saeidi et al., 2016; Barnes et al., 2022). In this work, we focus on the version of *aspect-based sentiment analysis* presented in the SemEval competitions (Pontiki et al., 2015, 2016), which includes several subtasks. Specifically, the tasks are aspect category detection (ACD), aspect term extraction (ATE), simultaneously detecting (aspect category, aspect term) tuples (ACTE), and detecting the sentiment polarity (APD)[1] of a given aspect term and category (see Figure 1 for examples).

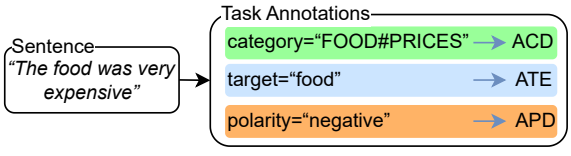

Figure 1: The example of the ABSA tasks.

In addition, we solve the target-aspect-sentiment detection task (TASD) (Wan et al., 2020), which

[1]The ACD, ATE, ACTE and APD tasks are named Slot1, Slot2, Slot1&2 and Slot3, respectively, in (Pontiki et al., 2015, 2016) under Subtask 1.

Proceedings of Recent Advances in Natural Language Processing, pages 1110–1120
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_118

aims to simultaneously detect the aspect category, aspect term and sentiment polarity.

This paper presents a novel approach for solving Czech sentiment classification and ABSA tasks using prompt-based fine-tuning. We utilize Czech monolingual BERT-like models and their language modelling ability to perform *prompting* for the APD and SC tasks. We use multilingual text-to-text generative models for the remaining ABSA tasks to generate textual predictions based on prompted input. Our approach enables us to solve all these ABSA tasks at once, and we show that it is superior to the traditional fine-tuning approach for them.

We also explore zero-shot and few-shot learning scenarios for APD and SC tasks and show that prompting leads to significantly better results with fewer training examples compared to traditional fine-tuning. Additionally, we demonstrate that pre-training on data from a target domain results in great improvements in a zero-shot scenario.

Our study provides pioneered results for prompt-based fine-tuning in Czech sentiment. Overall, our key contributions are the following: 1) We propose, to the best of our knowledge, the first prompt-based approach for sentiment analysis tasks in Czech. 2) We show the superior performance of our prompting approach over traditional fine-tuning for ABSA tasks. 3) We compare the two approaches and show that prompting achieves better results than traditional fine-tuning in few-shot scenarios.

2 Related Work

This section reviews prior works conducted on sentiment analysis in Czech. The prompt-based fine-tuning is a relatively new paradigm in NLP, and to the best of our knowledge, no research has yet explored its application on sentiment analysis in Czech. To partly address this research gap, we include prompt-based approaches for analogous sentiment analysis tasks in English.

2.1 Czech Sentiment Classification

The first approaches for sentiment analysis in Czech often utilized lexical features (Steinberger et al., 2011; Veselovská et al., 2012) and n-gram text representations in combination with classifiers like maximum entropy or Naive Bayes (Habernal et al., 2013). Subsequently, Brychcín and Habernal (2013) employed a mixture of supervised and unsupervised techniques to improve polarity detection in movie reviews. Similarly to Kim (2014), Lenc and Hercig (2016) used the convolutional neural network (CNN) and Long Short-Term Memory (LSTM) for SC of the same CSFD dataset we use in this work, see Section 3.1. The authors of (Libovický et al., 2018) added self-attention to an LSTM-based neural network and applied it to the CSFD dataset. A detailed survey of older approaches for Czech sentiment analysis is presented by Çano and Bojar (2019). In recent years, Czech Transformer-based models have been proposed and have shown great success in Czech sentiment analysis. Sido et al. (2021) introduced the first Czech BERT-like model, outperforming previous state-of-the-art (SotA) results in SC. Additionally, Straka et al. (2021) presented a pre-trained Czech version of the RoBERTa (Zhuang et al., 2021) model and demonstrated its effectiveness for the Czech language on Facebook posts. Přibáň and Steinberger (2021) provide the SotA results for three Czech polarity detection datasets. The most recent work comes from Přibáň et al. (2022); Přibáň and Steinberger (2022), where the authors investigate the possibility of performing zero-shot cross-lingual sentiment analysis and subjectivity classification between Czech and English with multilingual Transformer-based models like mBERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020).

2.2 Czech Aspect-Based Sentiment Analysis

The ABSA task in the Czech language has been much less studied in recent years and the existing approaches are usually outdated compared to recent sentiment classification methods. The pioneering research on Czech ABSA can be found in Steinberger et al. (2014), where the authors manually annotated and created a restaurant review dataset for the same tasks as in the SemEval 2014 competition (Pontiki et al., 2014). They provided results of baseline models based on Conditional Random Fields (CRF) and Maximum Entropy (ME) classifier. Tamchyna et al. (2015) built a dataset containing IT product reviews and provided baseline results with the CRF. Unlike in the mentioned Czech restaurant dataset, the IT product reviews are annotated with global sentiment and aspect terms but without any categorization and sentiment toward the terms. Hercig et al. (2016) extended the Czech restaurant review ABSA dataset and suggested several unsupervised methods to enhance the performance on ABSA tasks in Czech and English using the CRF and ME classifiers. They showed that

unsupervised methods can provide substantial improvements.

2.3 Prompt-Based and Related Approaches

As we already mentioned, there is no work for Czech sentiment analysis based on prompt-based fine-tuning. Therefore, we provide example studies focused on English sentiment analysis using prompt-based approaches or related methods.

Zhang et al. (2021b) formulate the ABSA tasks as a text generation problem. They propose two paradigms to deal with the ABSA tasks, namely annotation-style and extraction-style modelling, both generating textual output in a desired format. They utilize the English T5 (Raffel et al., 2020) text-to-text Transformer-based model and evaluate their approach on various ABSA tasks, including the TASD task, on datasets from the SemEval competitions (Pontiki et al., 2014, 2015, 2016). They showed the effectiveness of their approach by establishing new SotA results. Similarly, Zhang et al. (2021a) used the same English T5 model to solve a newly introduced ABSA task called *aspect sentiment quad prediction* by generating textual output. Another approach proposed by Gao et al. (2022) aims to solve multiple ABSA tasks at once. The authors applied the English T5 model to a prompt created from the individual ABSA tasks. They evaluated their model on the same datasets as Zhang et al. (2021b), outperforming the previously mentioned approach and achieving new SotA results.

Gao et al. (2021) experimented with prompt-based fine-tuning for SC. With the English T5 model, they automatically generated prompts for BERT and RoBERTa models, which they consequently fine-tuned for the SC task. They demonstrated that their few-shot approach leads to better results compared to traditional fine-tuning.

3 Data & Tasks Definition

In this section, we describe the aspect-based and sentiment classification datasets. Furthermore, we describe in more detail the ABSA tasks introduced in Section 1, on which this paper is focused.

3.1 Data for Sentiment Classification

For the SC task, where the goal is to assign one overall polarity label (*positive*, *negative* or *neutral*) for a given text, we employ the Czech CSFD dataset (Habernal et al., 2013). The dataset contains 91,381 movie reviews from the Czech movie database[2]. The reviews are annotated in a distant supervised way according to the star rating assigned to each review (0–1 stars as *negative*, 2–3 stars as *neutral*, 4–5 stars as *positive*). We use the training and testing split from Přibáň and Steinberger (2021), see Table 1.

Split	Positive	Negative	Neutral
train	24,573	23,840	24,691
test	6,324	5,876	6,077
total	30,897	29,716	30,768

Table 1: Statistics of the CSFD dataset.

For the additional pre-training (see Section 5.2), we downloaded 4.2M movie reviews (i.e. 1.8 GB of plain text) from the Czech movie database[2]. From the downloaded reviews, we removed all reviews present in the annotated CSFD dataset.

3.2 Data for Aspect-Based Sentiment Analysis

For the ABSA tasks, we use the Czech dataset (Hercig et al., 2016) from a restaurant domain which we convert into the SemEval 2016 competition (Pontiki et al., 2016) format to align with the ABSA tasks addressed in this paper. The dataset consists of 2,149 Czech restaurant reviews, which we split into the training and testing parts in a 75:25 ratio. The label distribution of the modified[3] ABSA dataset (Hercig et al., 2016) is shown in Table 2, along with the number of sentiment labels for aspect categories used in the APD and TASD tasks[4].

Split	Sentences	Positive	Negative	Neutral
train	1,612	1,231	1,197	336
test	537	420	426	61
total	2,149	1,651	1,623	397

Table 2: Statistics of the Czech ABSA dataset.

For the additional pre-training, we scraped 2.4M reviews of Czech restaurants from Google Maps[5], resulting in 330 MB of plain text. As restaurant reviews are shorter, the size is smaller than downloaded movie reviews. This resulted in 330 MB of plain text, a much smaller size compared to the downloaded movie reviews due to the shorter

[2] https://www.csfd.cz

[3] The dataset was converted into the SemEval 2016 competition (Pontiki et al., 2016).

[4] Because one review can contain multiple aspect categories, the number of sentiment labels does not sum up to the number of given sentences in Table 2.

[5] https://www.google.com/maps

length of restaurant reviews. We removed all reviews present in the annotated ABSA dataset.

3.3 Aspect-Based Sentiment Tasks Definition

Given the complexity and possible confusion in naming the aspect-based tasks we deal with in this paper, we briefly describe the tasks. As mentioned, in the ABSA tasks, we aim at the Czech restaurant reviews domain.

The ACD task aims to identify all *E#A* aspect categories towards which an opinion is expressed in a given sentence. The *E#A* represents a pair of one entity *E* (i.e. Ambience, Drinks, Food, Location, Restaurant and Service), and one attribute/aspect *A* (i.e. General, Miscellaneous, Prices, Quality, Style-Options). There are 14 predefined pairs of *E#A*, for example, *FOOD#PRICES*. Other than the predefined pair combinations are not allowed.

The ATE aims to extract the aspect term, i.e. the linguistic expression used in the given text that represents the entity *E* of each *E#A* pair. The aspect term does not have to be mentioned directly, for example, in the review: *"Expensive but delicious"*, the entity *E* is *Food*, but the aspect term is not present in the text. In such cases, the *NULL* value is assigned. The ACTE task focuses on extracting the aspect term and aspect category simultaneously.

The APD task's goal is to assign one of the three polarity labels (*positive*, *negative*, *neutral*) for all already identified (aspect category, aspect term) pairs in a given text. See Figure 1 for an example.

In the TASD task, the goal is to identify all (aspect category, aspect term, sentiment polarity) triplets simultaneously, which makes this task the most difficult task we solve.

4 Models & Approaches

We use pre-trained Transformer-based models as backbones for our experiments. We propose a method for solving multiple ABSA tasks concurrently with sequence-to-sequence models[6], which process text (sequence) as input and produce text (sequence) as output. We employ this approach for the ACD, ATE, ACTE and TASD subtasks. To the best of our knowledge, there are no Czech monolingual sequence-to-sequence models. Therefore, we use the large **mT5** (Xue et al., 2021) and large **mBART** (Tang et al., 2021) models, which are multilingual versions of the English T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) models, respectively.

We do not use these models for the APD task as they lack prior information about the aspect term and category, which they predict along with the sentiment. The APD task assumes that the model already knows the gold data for the aspect term and category, so we would have to modify the input and output format for the APD task to make a fair comparison. Changing the output format would also be required for the SC task.

Since we focus solely on the Czech language, we also wanted to evaluate Czech monolingual models. As stated above, there are no monolingual Czech sequence-to-sequence models, but only classical Czech monolingual BERT-like models such as **Czert** (Sido et al., 2021), **RobeCzech** (Straka et al., 2021) or **FERNET** (Lehečka and Švec, 2021). Unfortunately, these models are unsuitable for our proposed approach, so we use them only for the APD and SC tasks. These models consist only of the encoder part of the Transformer architecture.

4.1 Sequence-to-Sequence Models

We employ the multilingual sequence-to-sequence models (mT5, mBART) to solve several ABSA tasks at once. These models consist of two parts of the Transformer architecture: the *encoder* and the *decoder*. Given the input sequence x, the encoder transforms it into a contextualized sequence $\mathbf{e}$. The decoder then models the conditional probability distribution of the target sequence y given the encoded input $\mathbf{e}$ as $P_{\boldsymbol{\Theta}}(y|\mathbf{e})$, where $\boldsymbol{\Theta}$ are the parameters of the model. At each step, i, the decoder output y_i is computed based on the previous outputs $y_0, \ldots, y_{i-1}$ and the encoded input $\mathbf{e}$. During fine-tuning, we update all model parameters.

4.1.1 Traditional Fine-Tuning

Because the output of sequence-to-sequence models is text, we have to convert our discrete ABSA labels to the textual format inspired by Zhang et al. (2021a). For each example in the ABSA dataset, we construct the label as "*c is $P_p(p)$, given the expression: a*", where c is the aspect category, a the aspect term and $P_p(p)$ a mapping function that maps the sentiment polarity p as

$$P_p(p) = \begin{cases} great & \text{if } p \text{ is } positive, \\ ok & \text{if } p \text{ is } neutral, \\ bad & \text{if } p \text{ is } negative. \end{cases} \quad (1)$$

[6] Also known as *text-to-text* models.

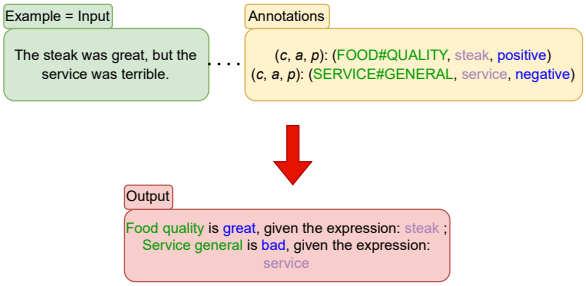

Figure 2: Example of the input and output construction for the T5 model with traditional fine-tuning.

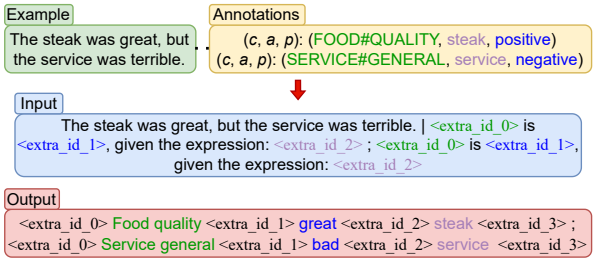

Figure 3: Example of the input and output construction for the T5 model with prompting.

For example, given the review: "*The steak was very tasty*" the following label is generated: "*Food quality* is *great*, given the expression: *steak*". If an example has multiple annotation triplets[7], we concatenate the labels with semicolons.

In this scenario, the model's input is the text (review), and the expected output is the textual label. The model's parameters are optimized to produce textual label in the desired format. Figure 2 shows an example of creating the input and target for the mT5 model with traditional fine-tuning.

4.1.2 Prompt-Based Fine-Tuning

For the prompt-based method, we expand the input review x with a template t to create a final input x': $x' = x + | + t$. The template has the same form as the label in the traditional fine-tuning method. The number of transformed triplets in the prompt corresponds to the number of triplets provided for one example. We design the prompt for the mT5 and mBART models differently because their training objectives differ.

The mT5 model aims to reconstruct randomly selected continuous spans of input text that are masked by sentinel tokens `<extra_id_id>` during pre-training. Here, *id* refers to the ID of the sentinel token, which starts from zero and increments by one. The model replaces non-masked spans of text with sentinel tokens. In our method, we replace the aspect category with `<extra_id_0>`, the sentiment polarity with `<extra_id_1>`, and the aspect term with `<extra_id_2>` to create the final input, which is inspired by work in (Gao et al., 2022). The output of the mT5 model consists of the aspect category, sentiment polarity and aspect term separated by sentinel tokens. Figure 3 shows an example of creating the input and target for the mT5 model with prompting.

[7]Each review can have multiple aspect categories and aspect terms, thus multiple triplet annotations.

Unlike T5, the BART model reconstructs the entire input text rather than just masked spans. Furthermore, the BART model utilizes the `<mask>` token instead of sentinel tokens.

4.1.3 Task Predictions

As mentioned earlier, we use sequence-to-sequence models to solve multiple ABSA tasks simultaneously, namely the ACD, ATE, ACTE and TASD. Each task aims to predict different components of the annotation triplet (aspect category, aspect term, sentiment polarity). We generate one output for all tasks and use only the relevant part of the output for each task while discarding the rest. We can extract the relevant part for each task because the model is trained to generate output in the expected format. For instance, we extract only the aspect term from the generated output in the ATE task. For the ATE task, we consider only distinct targets and discard *NULL* targets for the evaluation. For the ACD, ACTE and TASD tasks, we ignore duplicate occurrences of the predicted targets (e.g. aspect category for the ACD task).

4.2 Sentiment Polarity Classification Models

We use Czech BERT-like (encoder-based) models (i.e. Czert, RobeCzech, FERNET) to classify the sentiment polarity. These models convert an input sequence $x = w_1, \ldots, w_k$ of k tokens into a sequence of hidden vectors $\mathbf{h} = \mathbf{h}_0, \mathbf{h}_1, \ldots, \mathbf{h}_k$. For the APD task, we create n input-target pairs for each example, where n is the number of annotation triplets for that example.

4.2.1 Traditional Fine-Tuning

We employ a linear layer on top of the model to make a prediction. It computes the probability of a label y from a label space $\mathcal{Y} \in \{positive, negative, neutral\}$ for the input x_i as

$$P_{\Theta}(y|x_i) = \text{softmax}(\mathbf{W}\mathbf{h}_{\texttt{[CLS]}} + b), \quad (2)$$

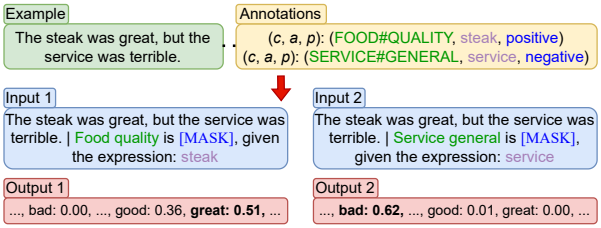

Figure 4: Example of the input and output construction for the classification model using prompting.

where $\boldsymbol{\Theta}$ denotes all the parameters to be fine-tuned, including task-specific ones ($\mathbf{W}$ and b). The hidden vector $\mathbf{h}_{\texttt{[CLS]}}$ represents the artificial classification `[CLS]` token corresponding to the first hidden vector of the input sequence, i.e. $\mathbf{h}_{\texttt{[CLS]}} = \mathbf{h}_0$, and represents the entire input sequence.

In the case of the ABSA dataset and the SC of the aspect term and category, we append the aspect term and category to the beginning of the input so that the model has the knowledge of the specific tuple by which to make predictions.

4.2.2 Prompt-Based Fine-Tuning

For prompt-based fine-tuning, we exploit the fact that the models were pre-trained by the masked language modelling task (Devlin et al., 2019). We use the language modelling property of the model to generate a token that represents the polarity label.

During prompt-based fine-tuning, we create a new input x' from the original input x by appending a task-specific prompt. The prompt has one answer slot represented by a `[MASK]` token, which the model fills with the highest-probability token from its vocabulary for the given context. Each label from label space $\mathcal{Y}$ is mapped to a word from the model's vocabulary $\mathcal{V}$ using a mapping $\mathcal{M} = \mathcal{Y} \rightarrow \mathcal{V}$, which is for Czech defined as follows

$$P_p(p) = \begin{cases} \textit{dobrý} & \text{if } p \text{ is } \textit{positive}, \\ \textit{ok} & \text{if } p \text{ is } \textit{neutral}, \\ \textit{špatný} & \text{if } p \text{ is } \textit{negative}. \end{cases} \quad (3)$$

Figure 4 shows an example of input construction with desired outputs for the ABSA task.

We trim the original input of long reviews before appending the prompt to ensure that the new input x' fits into the model. We use different prompts for each dataset. For the CSFD dataset, we use the prompt "*Je to* `[MASK]` *film*" ("*It is a* `[MASK]` *movie*" in English). For the ABSA dataset, the prompt is structured as "*c je* `[MASK]`, *dáno výrazem: a*" ("*c is* `[MASK]`, *given the expression: a*" in English), where c is the aspect category (translated to Czech) and a is the aspect term.

5 Experiments & Results

In our experiments, we fine-tune the sequence-to-sequence models (mBART, mT5) for the ATE, ACD, ACTE, and TASD tasks on the entire ABSA dataset using both traditional and prompt-based fine-tuning approaches and we report the results as micro F1 scores. The BERT-like (encoder-based) models (Czert, RobeCzech and FERNET) are fine-tuned for the APD and SC tasks and report results as accuracy. For these tasks, we further experiment with zero-shot and few-shot scenarios, as well as additional pre-training of the Czech models.

To ensure the reliability of our results, we perform each experiment five times with different random seed initialization and report the average scores along with a 95% confidence interval. We provide the training details in Appendix A.1.

5.1 Few-Shot and Zero-Shot Setting

In the few-shot setting, we fine-tune the models on the first n examples of the training data using a fixed training set to ensure a fair comparison between models, as recommended by Schick and Schütze (2021). In the zero-shot setting, models are evaluated on the test set without any fine-tuning.

5.2 Additional Pre-Training

For the APD and SC tasks, we were interested in whether additional pre-training in the task domain helps to improve results. Therefore, we further pre-train the three Czech models (Czert, RobeCzech and FERNET) with the masked language modelling task on restaurant reviews and movie reviews for the APD and SC tasks, respectively. See Appendix A.2 for details.

5.3 Results for Aspect-Based Sentiment

Table 3 shows the results achieved by the sequence-to-sequence models. The prompting approach (PT-FT) significantly enhances the performance of both models. Without prompting, i.e. with the traditional fine-tuning (TR-FT), mBART performs better than mT5. However, with prompting, mT5 performs better than or similar to mBART.

The best results are achieved on the ACD task. For this task, there is a predefined set of categories. In contrast, the ATE task poses a greater challenge,

Model	Approach	Task			
		ACD	ATE	ACTE	TASD
mT5	TR-FT	$75.5^{\pm 1.8}$	$66.5^{\pm 2.5}$	$56.4^{\pm 1.0}$	$48.0^{\pm 1.0}$
	PT-FT	**85.5**$^{\pm 1.2}$	**84.8**$^{\pm 1.6}$	**75.0**$^{\pm 1.9}$	**67.3**$^{\pm 1.7}$
mBART	TR-FT	$78.7^{\pm 1.6}$	$78.9^{\pm 1.3}$	$67.2^{\pm 1.4}$	$57.5^{\pm 1.7}$
	PT-FT	$83.3^{\pm 0.7}$	$83.4^{\pm 0.6}$	$71.9^{\pm 1.6}$	$61.7^{\pm 0.7}$

Table 3: Results of the sequence-to-sequence models as micro F1 scores on different ABSA tasks with traditional fine-tuning (TR-FT) and prompt-based fine-tuning (PT-FT). The best results for each task are in **bold**. Underlined results indicate significantly better performance between the two fine-tuning styles for a given model and task.

as the extracted term can be an arbitrarily long sequence of different words, making this task more difficult. The ACTE is even more challenging since the model has to simultaneously predict the aspect term and category. The TASD task is the most difficult of the solved tasks because the model must predict the aspect term, aspect category and sentiment polarity simultaneously. Since our study is the first to focus on these tasks in the Czech language, we lack a basis for comparison with other studies.

Table 4 shows the results of the APD task. Traditional fine-tuning performs significantly better than prompting in the zero-shot setting. Prompting outperforms traditional fine-tuning when using a small number of examples for training. In the rest of the results, both fine-tuning approaches perform similarly. The domain pre-training improves the results of all models, especially for traditional fine-tuning.

5.4 Sentiment Classification Results

Table 5 shows the sentiment classification results on the CSFD dataset, along with the current SotA results. In the zero-shot setting, the traditional fine-tuning approach (TR-FT) yields random results around 35–38%. This is expected because the linear layer[8] on top of the model is not trained and the CSFD dataset contains three roughly balanced classes. On the other hand, the zero-shot scenario with the prompt-based approach[9] (PT-FT) combined with the additional domain pre-training provides significantly better results for Czert and FERNET models, achieving 48.2% and 59.2%, respectively.

We observed that prompting consistently outperforms traditional fine-tuning in the few-shot scenario with 10 and 20 training examples. In contrast, traditional fine-tuning yields better results when using 100, 500 and 1,000 examples. Results are comparable for both approaches when the model is trained on all examples and 50 examples. Domain pre-training improves the results in most cases, especially when using only a small number of examples. Notably, the FERNET model achieved the best result of 88.2% accuracy, surpassing the current SotA by 2.8%.

5.5 Discussion

The prompt designed for the APD task might be more suitable than the prompt for the SC task, which may explain why prompting is worse only in one case than traditional fine-tuning outside of the zero-shot setting, while traditional fine-tuning outperforms prompting more often in the SC task.

The reason why the sequence-to-sequence models perform better with prompting than with traditional fine-tuning may be that the prompting matches these models' pre-training objectives closely. Additionally, these models possess some prior information about the number of sentiment triplets they should generate in the prompt, which the traditional fine-tuned models do not.

Our research indicates that the sequence-to-sequence models have no problems generating the output in the required format, which is crucial to extract the targets. However, when using traditional fine-tuning, the mT5 model occasionally generates repeated transformed triplets and lacks diversity in its output more frequently than the mBART model, which may explain why the mBART model outperforms the mT5 model with traditional fine-tuning.

We observe a common trend in results for SC and APD tasks, whereby the prompting approach with a smaller number of training examples outperforms the traditional fine-tuning, which is consistent with conclusions from Gao et al. (2021).

For prompting in few-shot and zero-shot scenarios, a mapping function that maps one sentiment to multiple words instead of one specific word would likely lead to better results, which can be explored in future work.

[8]The layer always returns one of three possible labels, thus if the dataset is perfectly balanced, the random (and also lowest) expected accuracy is $33.\overline{3}\%$.

[9]In this case, the model can predict any word from the model vocabulary $\mathcal{V}$; therefore, the potential lowest expected random accuracy is close to zero $(1/|\mathcal{V}|)$.

	Czert		RobeCzech		FERNET	
	TR-FT original/pre-train	PT-FT original/pre-train	TR-FT original/pre-train	PT-FT original/pre-train	TR-FT original/pre-train	PT-FT original/pre-train
Zero-shot						
	47.1$^{\pm 0.7}$/42.4$^{\pm 6.4}$	5.3$^{\pm 0.0}$/5.3$^{\pm 0.0}$	**47.4**$^{\pm 2.5}$/42.3$^{\pm 3.5}$	8.4$^{\pm 0.0}$/3.8$^{\pm 0.0}$	43.6$^{\pm 2.4}$/43.2$^{\pm 3.3}$	0.8$^{\pm 0.0}$/3.2$^{\pm 0.0}$
Fine-tuning (few-shot)						
10	46.0$^{\pm 1.9}$/55.5$^{\pm 3.9}$	67.6$^{\pm 3.6}$/77.5$^{\pm 5.0}$	47.5$^{\pm 3.0}$/65.5$^{\pm 6.9}$	77.3$^{\pm 3.4}$/**81.9**$^{\pm 1.4}$	48.8$^{\pm 2.0}$/66.3$^{\pm 4.0}$	77.6$^{\pm 6.5}$/76.9$^{\pm 3.7}$
20	54.6$^{\pm 6.0}$/76.5$^{\pm 5.4}$	74.3$^{\pm 1.3}$/80.4$^{\pm 1.1}$	59.7$^{\pm 2.0}$/63.4$^{\pm 7.3}$	78.5$^{\pm 2.0}$/**82.8**$^{\pm 1.1}$	62.6$^{\pm 1.6}$/79.4$^{\pm 4.2}$	72.7$^{\pm 3.0}$/78.5$^{\pm 2.1}$
50	66.0$^{\pm 4.6}$/83.4$^{\pm 2.3}$	75.2$^{\pm 2.0}$/80.9$^{\pm 2.6}$	75.3$^{\pm 2.5}$/**86.7**$^{\pm 1.4}$	83.0$^{\pm 1.7}$/85.7$^{\pm 0.6}$	71.9$^{\pm 2.1}$/83.7$^{\pm 3.3}$	84.5$^{\pm 1.4}$/86.6$^{\pm 1.4}$
100	66.6$^{\pm 3.0}$/80.4$^{\pm 1.4}$	75.9$^{\pm 0.7}$/81.2$^{\pm 1.8}$	76.3$^{\pm 6.9}$/84.3$^{\pm 1.6}$	83.3$^{\pm 1.3}$/**85.5**$^{\pm 1.0}$	71.6$^{\pm 2.7}$/82.5$^{\pm 2.1}$	84.1$^{\pm 1.6}$/85.1$^{\pm 1.7}$
500	81.4$^{\pm 2.1}$/84.1$^{\pm 1.4}$	82.6$^{\pm 1.0}$/84.3$^{\pm 0.9}$	84.0$^{\pm 1.4}$/86.6$^{\pm 0.3}$	85.6$^{\pm 1.8}$/85.3$^{\pm 0.8}$	84.5$^{\pm 1.1}$/83.8$^{\pm 0.5}$	84.2$^{\pm 1.1}$/**86.7**$^{\pm 1.6}$
1,000	82.0$^{\pm 1.1}$/83.4$^{\pm 1.6}$	82.7$^{\pm 1.0}$/83.2$^{\pm 1.5}$	83.1$^{\pm 2.7}$/**87.4**$^{\pm 2.1}$	85.3$^{\pm 1.7}$/87.2$^{\pm 1.5}$	84.6$^{\pm 0.8}$/87.0$^{\pm 1.1}$	85.9$^{\pm 0.5}$/85.9$^{\pm 0.7}$
Fine-tuning (full)						
	83.2$^{\pm 1.4}$/85.0$^{\pm 1.1}$	84.2$^{\pm 1.1}$/87.0$^{\pm 1.3}$	85.2$^{\pm 1.6}$/88.4$^{\pm 0.9}$	87.3$^{\pm 1.4}$/**88.7**$^{\pm 1.0}$	86.0$^{\pm 0.4}$/88.4$^{\pm 0.7}$	87.5$^{\pm 1.2}$/88.5$^{\pm 0.7}$

Table 4: Results for the ABSA dataset on APD task as accuracy with prompt-based fine-tuning (PT-FT) and traditional fine-tuning (TR-FT) approaches. The best results for a given configuration are in **bold**. Underlined results indicate significantly better performance between the two fine-tuning styles for a given model (both original and with additional pre-training) and the number of training examples.

	Czert		RobeCzech		FERNET	
	TR-FT original/pre-train	PT-FT original/pre-train	TR-FT original/pre-train	PT-FT original/pre-train	TR-FT original/pre-train	PT-FT original/pre-train
Zero-shot						
	35.0$^{\pm 0.7}$/35.7$^{\pm 2.2}$	11.8$^{\pm 0.0}$/48.2$^{\pm 0.0}$	36.3$^{\pm 2.9}$/35.7$^{\pm 5.0}$	12.7$^{\pm 0.0}$/8.9$^{\pm 0.0}$	38.2$^{\pm 1.1}$/36.8$^{\pm 4.0}$	5.8$^{\pm 0.0}$/**59.2**$^{\pm 0.0}$
Fine-tuning (few-shot)						
10	43.4$^{\pm 1.9}$/54.6$^{\pm 2.0}$	50.3$^{\pm 0.6}$/60.4$^{\pm 0.8}$	46.2$^{\pm 3.0}$/61.3$^{\pm 0.4}$	54.6$^{\pm 1.4}$/**62.4**$^{\pm 1.5}$	48.8$^{\pm 2.5}$/55.1$^{\pm 3.3}$	56.4$^{\pm 0.6}$/61.5$^{\pm 0.5}$
20	47.4$^{\pm 3.1}$/60.9$^{\pm 3.6}$	51.5$^{\pm 0.3}$/65.2$^{\pm 1.0}$	48.4$^{\pm 3.3}$/65.4$^{\pm 4.1}$	56.0$^{\pm 0.9}$/**72.3**$^{\pm 0.8}$	57.8$^{\pm 2.9}$/62.6$^{\pm 2.8}$	62.6$^{\pm 1.7}$/67.5$^{\pm 0.3}$
50	57.1$^{\pm 3.6}$/71.0$^{\pm 1.2}$	58.7$^{\pm 0.8}$/70.9$^{\pm 0.7}$	56.7$^{\pm 4.7}$/**78.5**$^{\pm 0.9}$	60.3$^{\pm 2.2}$/77.1$^{\pm 0.4}$	66.6$^{\pm 2.4}$/74.7$^{\pm 4.2}$	67.4$^{\pm 1.8}$/75.7$^{\pm 3.9}$
100	64.3$^{\pm 0.8}$/73.9$^{\pm 0.7}$	61.6$^{\pm 0.6}$/72.8$^{\pm 0.2}$	69.8$^{\pm 1.1}$/**80.1**$^{\pm 0.3}$	67.7$^{\pm 0.9}$/78.6$^{\pm 0.2}$	74.1$^{\pm 0.4}$/79.8$^{\pm 1.0}$	72.1$^{\pm 0.4}$/78.2$^{\pm 1.2}$
500	70.7$^{\pm 0.2}$/75.7$^{\pm 1.0}$	69.2$^{\pm 0.4}$/75.8$^{\pm 0.3}$	74.3$^{\pm 0.7}$/82.2$^{\pm 0.2}$	73.9$^{\pm 0.5}$/81.1$^{\pm 0.2}$	77.3$^{\pm 0.3}$/**82.5**$^{\pm 0.5}$	76.4$^{\pm 0.1}$/81.8$^{\pm 0.4}$
1,000	72.7$^{\pm 0.1}$/76.6$^{\pm 0.1}$	71.2$^{\pm 0.2}$/76.1$^{\pm 0.9}$	76.2$^{\pm 0.8}$/82.7$^{\pm 0.2}$	75.7$^{\pm 0.7}$/82.3$^{\pm 0.1}$	78.4$^{\pm 0.3}$/**83.0**$^{\pm 0.8}$	77.6$^{\pm 0.3}$/82.3$^{\pm 0.4}$
Fine-tuning (full)						
	85.3$^{\pm 0.1}$/86.5$^{\pm 0.1}$	85.3$^{\pm 0.1}$/86.3$^{\pm 0.1}$	87.1$^{\pm 0.0}$/88.0$^{\pm 0.1}$	87.0$^{\pm 0.3}$/87.9$^{\pm 0.2}$	87.3$^{\pm 0.1}$/**88.2**$^{\pm 0.1}$	87.2$^{\pm 0.2}$/87.7$^{\pm 0.7}$
Přibáň and Steinberger (2021)	84.8$^{\pm 0.1}$/ -	-	-	-	-	
Lehečka and Švec (2021)	-	-	85.0$^{\pm 0.4}$/ -	-	85.4$^{\pm 0.3}$/ -	-

Table 5: Sentiment classification results for the CSFD dataset as accuracy with prompt-based fine-tuning (PT-FT) and traditional fine-tuning (TR-FT) approaches. The best results for a given configuration are in **bold**. Underlined results indicate significantly better performance between the two fine-tuning styles for a given model (both original and with additional pre-training) and the number of training examples.

6 Conclusion

In this work, we introduced a sequence-to-sequence method that solves multiple ABSA tasks simultaneously and can be used with both traditional fine-tuning and prompting. Experiments on the Czech dataset show that prompting significantly improves performance. Furthermore, we proposed a method for sentiment classification that can also be used with prompting and traditional fine-tuning. We evaluate this method on two Czech datasets with three monolingual Czech models and demonstrate the effectiveness of prompting for few-shot fine-tuning, where prompting consistently outperforms the traditional approach. Finally, we show that pre-training on the domain data significantly enhances the results, especially in a zero-shot scenario.

Acknowledgments

This work has been partly supported by grant No. SGS-2022-016 Advanced methods of data processing and analysis. Computational resources were provided by the e-INFRA CZ project (ID:90140), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval 2022 task

10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States. Association for Computational Linguistics.

Tomáš Brychcín and Ivan Habernal. 2013. Unsupervised improving of sentiment analysis using global target context. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 122–128, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Computational Linguistics.

Tianhao Gao, Jun Fang, Hanyu Liu, Zhiyuan Liu, Chao Liu, Pengzhang Liu, Yongjun Bao, and Weipeng Yan. 2022. LEGO-ABSA: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7002–7012, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Ivan Habernal, Tomáš Ptáček, and Josef Steinberger. 2013. Sentiment analysis in Czech social media using supervised machine learning. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 65–74, Atlanta, Georgia. Association for Computational Linguistics.

Tomáš Hercig, Tomáš Brychcín, Lukáš Svoboda, Michal Konkol, and Josef Steinberger. 2016. Unsupervised methods to improve aspect-based sentiment analysis in czech. *Computación y Sistemas*, 20(3):365–375.

Ehsan Hosseini-Asl, Wenhao Liu, and Caiming Xiong. 2022. A generative language model for few-shot aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 770–787, Seattle, United States. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Jan Lehečka and Jan Švec. 2021. Comparison of czech transformers on text classification tasks. In *Statistical Language and Speech Processing*, pages 27–37, Cham. Springer International Publishing.

Ladislav Lenc and Tomás Hercig. 2016. Neural networks for sentiment analysis in czech. In *Proceedings of the 16th ITAT: Slovenskočeský NLP workshop (SloNLP 2016)*, volume 1649 of *CEUR Workshop Proceedings*, pages 48–55, Bratislava, Slovakia. Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, CreateSpace Independent Publishing Platform.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jindrich Libovický, Rudolf Rosa, Jindrich Helcl, and Martin Popel. 2018. Solving three czech nlp tasks with end-to-end neural models. In *ITAT*, pages 138–143.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Pavel Přibáň, Jakub Šmíd, Adam Mištera, and Pavel Král. 2022. Linear transformations for cross-lingual sentiment analysis. In *Text, Speech, and Dialogue*, pages 125–137, Cham. Springer International Publishing.

Pavel Přibáň and Josef Steinberger. 2021. Are the multilingual models better? improving Czech sentiment with transformers. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1138–1149, Held Online. INCOMA Ltd.

Pavel Přibáň and Josef Steinberger. 2022. Czech dataset for cross-lingual subjectivity classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1381–1391, Marseille, France. European Language Resources Association.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. 2016. SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1546–1556, Osaka, Japan. The COLING 2016 Organizing Committee.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4603–4611. PMLR.

Jakub Sido, Ondřej Pražák, Pavel Přibáň, Jan Pašek, Michal Seják, and Miloslav Konopík. 2021. Czert – Czech BERT-like model for language representation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1326–1338, Held Online. INCOMA Ltd.

Josef Steinberger, Tomáš Brychcín, and Michal Konkol. 2014. Aspect-level sentiment analysis in Czech. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–30, Baltimore, Maryland. Association for Computational Linguistics.

Josef Steinberger, Polina Lenkova, Mohamed Ebrahim, Maud Ehrmann, Ali Hurriyetoglu, Mijail Kabadjov, Ralf Steinberger, Hristo Tanev, Vanni Zavarella, and Silvia Vázquez. 2011. Creating sentiment dictionaries via triangulation. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 28–36, Portland, Oregon. Association for Computational Linguistics.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. RobeCzech: Czech RoBERTa, a Monolingual Contextualized Language Representation Model. In *Text, Speech, and Dialogue*, pages 197–209, Cham. Springer International Publishing.

Ales Tamchyna, Ondrej Fiala, and Katerina Veselovská. 2015. Czech aspect-based sentiment analysis: A new dataset and preliminary results. In *ITAT*, pages 95–99.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Katerina Veselovská, Jan Hajic, and Jana Sindlerová. 2012. Creating annotated resources for polarity classification in czech. In *KONVENS*, pages 296–304.

Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z. Pan. 2020. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9122–9129.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Erion Çano and Ondřej Bojar. 2019. Sentiment analysis of czech texts: An algorithmic survey. In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence*, pages 973–979, Setúbal, Portugal. SCITEPRESS Digital Library.

A Appendix

A.1 Hyper-parameters & Training Details

We train the models with different hyper-parameters and select the best-performing model based on the performance on the validation data, including the number of epochs. The CSFD dataset is already split into training and validation data. For the ABSA dataset, we use 10% of the training data as validation data. The final experiments are conducted on all training data and evaluated on the test data.

We use a batch size of 64 and train the sequence-to-sequence models for up to 35 epochs. For the mT5 model, we search for a learning rate from {1e-4, 3e-4}, while for the mBART model, we search for a learning rate from {5e-5, 1e-5}. We use greedy search for simplicity because experiments with a beam search with beam sizes 3 and 5 lead to similar performance.

For the models for sentiment polarity classification, we search for a learning rate from {5e-5, 1e-5}. We use up to 10 epochs and a batch size of 16 for the CSFD dataset and up to 50 epochs and a batch size of 64 for the ABSA dataset.

We optimize the cross-entropy loss for all the models. All the models have the maximum input sequence length limited to 512 tokens. We use the AdaFactor (Shazeer and Stern, 2018) optimizer for the mT5 model and AdamW (Loshchilov and Hutter, 2019) for the rest of the models. We keep the default dropout value for all the models, which is 0.1.

For text generation with the sequence-to-sequence models, we use the *AutoModelForSeq2SeqLM* class with greedy search decoding from the HuggingFace library[10]. We tried different configurations of the beam search decoding algorithm (Freitag and Al-Onaizan, 2017), but it provides the same results as the greedy search algorithm, so we employ the greedy search algorithm for simplicity.

A.2 Details of Additional Pre-Training

The additional pre-training of Czert, RobeCzech and FERNET models on data from a specific task domain (restaurant reviews and movie reviews) is performed with the masked language modelling task (Devlin et al., 2019). The pre-training process was carried out with a batch size of 512 and a maximum input sequence length of 512 for all models. We optimize the models with the cross-entropy loss function and AdamW (Loshchilov and Hutter, 2019) optimizer for 20K batches (steps). We use a learning rate of 5e-5 with linear decay. The word masking probability is set to 15%.

[10] https://huggingface.co