Czech Dataset for Complex Aspect-Based Sentiment Analysis Tasks

Jakub Šmíd[*], Pavel Přibáň[*], Ondřej Pražák[*, †], Pavel Král[*, †]
University of West Bohemia, Faculty of Applied Sciences,
[*]Department of Computer Science and Engineering,
[†]NTIS – New Technologies for the Information Society
Univerzitní 2732/8, 301 00 Pilsen, Czech Republic
{jaksmid, pribanp, ondfa, pkral}@kiv.zcu.cz
https://nlp.kiv.zcu.cz

Abstract

In this paper, we introduce a novel Czech dataset for aspect-based sentiment analysis (ABSA), which consists of 3.1K manually annotated reviews from the restaurant domain. The dataset is built upon the older Czech dataset, which contained only separate labels for the basic ABSA tasks such as aspect term extraction or aspect polarity detection. Unlike its predecessor, our new dataset is specifically designed for more complex tasks, e.g. target-aspect-category detection. These advanced tasks require a unified annotation format, seamlessly linking sentiment elements (labels) together. Our dataset follows the format of the well-known SemEval-2016 datasets. This design choice allows effortless application and evaluation in cross-lingual scenarios, ultimately fostering cross-language comparisons with equivalent counterpart datasets in other languages. The annotation process engaged two trained annotators, yielding an impressive inter-annotator agreement rate of approximately 90%. Additionally, we provide 24M reviews without annotations suitable for unsupervised learning. We present robust monolingual baseline results achieved with various Transformer-based models and insightful error analysis to supplement our contributions. Our code and dataset are freely available for non-commercial research purposes.

Keywords: Aspect-Based Sentiment Analysis, Dataset Construction, Czech

1. Introduction

Sentiment analysis (SA) is a widely recognized and fundamental field of natural language processing that aims to understand and identify subjective information in text (Liu, 2012). Sentiment classification (SC), known as polarity detection, is a common task within sentiment analysis that aims to classify a given text into one of pre-defined categories, such as *positive*, *negative* or *neutral*.

Aspect-based sentiment analysis (ABSA) is a more fine-grained task than SC. ABSA focuses on extracting detailed information about entities, their aspects and opinions expressed regarding these aspects. ABSA generally aims to identify the sentiment associated with each aspect or characteristic of a product or service. For instance, in restaurant reviews, opinions are not limited to the overall food quality but extend to other aspects like service, location, and atmosphere. ABSA includes sentiment elements (Zhang et al., 2022), such as aspect term (a), aspect category (c), and sentiment polarity (p). In the review (s): *"Delicious steak"*, these elements are *"steak"*, *food quality*, and *positive*, respectively.

ABSA involves several tasks (Zhang et al., 2022). Initially, research focused on identifying each sentiment element separately, such as aspect term extraction (ATE) or aspect category detection (ACD) (Pontiki et al., 2014). Recently, the focus has shifted to tasks that require linking sentiment elements in annotations, such as aspect polarity detection (APD). This linking also allows to predict more sentiment elements together, such as aspect-category-term extraction (ACTE) (Pontiki et al., 2015), unified end-to-end ABSA (E2E-ABSA) (Wang et al., 2018), and target-aspect-category detection (TASD) (Wan et al., 2020). Table 1 shows input and output examples for selected ABSA tasks.

Task	Input	Output	Example output
ATE	s	$\{a\}$	{"steak", "water"}
ACD	s	$\{c\}$	{food, drinks}
APD	s, "steak", food	p	POS
E2E-ABSA	s	$\{(a, p)\}$	{("steak", POS), ("water", NEG)}
ACTE	s	$\{(a, c)\}$	{("steak", food), ("water", drinks)}
TASD	s	$\{(a, c, p)\}$	{("steak", food, POS), ("water", drinks, NEG)}

Table 1: Input and output format for ABSA tasks for a review s: *"Delicious steak but expensive water"*.

For ABSA, several datasets have been built over time, including SemEval-2014 (Pontiki et al., 2014), SemEval-2015 (Pontiki et al., 2015) and SemEval-2016 (Pontiki et al., 2016) datasets, SentiHood (Saeidi et al., 2016) or Japanese dataset introduced by Nakayama et al. (2022). The datasets are mainly created for the English language except for the SemEval-2016, which also contains Arabic, Chinese, Dutch, French, Russian, Spanish and Turkish annotations. Fan et al. (2019) and Zhang et al. (2021a) provide datasets with opinion terms annotations. Steinberger et al. (2014) and Hercig et al. (2016) introduced Czech datasets in the same format as the SemEval-2014 dataset, allowing a separate evaluation of the ATE, ACD and sentiment

LREC-COLING 2024, pages 4299–4310
20-25 May, 2024.

classification tasks in the Czech language. Similarly, Tamchyna et al. (2015) presented a dataset with IT product reviews but annotated it only with a global review sentiment and aspect terms.

Unfortunately, the mentioned Czech datasets do not link the aspect term and category annotations, making it impossible to solve tasks where these two sentiment elements are predicted together, namely the TASD, ACTE and APD tasks. Therefore, the primary motivation of this paper is to provide a dataset that enables the evaluation of these advanced tasks in Czech. Consequently, the new dataset will allow cross-language comparison.

This paper presents a new dataset of 3,189 restaurant reviews tailored for complex ABSA tasks such as TASD and ACTE, which require annotations in a unified format linking individual labels together. Additionally, we crawled a set of 24M raw reviews intended for unsupervised learning. We re-annotated the existing Czech dataset (Hercig et al., 2016) and expanded it with more than 1,000 new reviews. The dataset adheres to the SemEval-2016 format, allowing evaluation of the more complex tasks and well-established existing tasks of ABSA in Czech. We describe the process of dataset creation and annotation. Two trained annotators annotated the dataset with an inter-annotator agreement of approximately 90%.

We conduct a series of experiments and present robust baseline results utilizing Transformer-based models for the older ATE and ACD tasks, achieving 83.5% and 85.7% of the F1-score, respectively. Furthermore, we report baseline results for the complex APD, E2E, ACTE, and TASD tasks with 91.4%, 75.5%, 67.3%, and 59.3%, respectively. Finally, we provide an error analysis of sequence-to-sequence models, showing their weaknesses and limitations.

Our main contributions are the following: 1) We introduce a new Czech ABSA dataset[1] in the restaurant domain that allows solving more complex ABSA tasks and cross-lingual comparisons with SemEval-2016 datasets. 2) We perform experiments with Transformer-based models and provide robust baseline results with an error analysis.

2. Related Work

This section is devoted to existing and, according to our judgment, important ABSA datasets. Further, we review prior and recent works focused on aspect-based sentiment analysis, especially in Czech.

[1]The annotated dataset, including its training splits, and code are freely available for research purposes at https://github.com/biba10/Czech-Dataset-for-Complex-ABSA-Tasks.

2.1. ABSA Datasets

Several ABSA datasets have been proposed. The SemEval-2014 dataset (Pontiki et al., 2014) contains English reviews from restaurants and laptop domains. The SemEval-2015 dataset (Pontiki et al., 2015) is based on the SemEval-2014 dataset with a more unified annotation format that links sentiment elements together. The SemEval-2016 dataset (Pontiki et al., 2016) is extended to new domains and provides more languages besides English, namely Arabic, Chinese, Dutch, French, Russian, Spanish and Turkish. These datasets lack the annotation of opinion terms. Fan et al. (2019) provide the dataset with annotated opinion terms for English. Zhang et al. (2021a) introduce English datasets for different domains containing the annotations of four sentiment elements. The MAMS dataset (Jiang et al., 2019) is another dataset that focuses on the restaurant domain. Twitter is another valuable linguistic data resource, and Dong et al. (2014) constructed a dataset from Twitter comments. The SentiHood dataset (Saeidi et al., 2016) is derived from a question-answering platform where users discuss urban neighbourhoods. Nakayama et al. (2022) introduced the ABSA dataset for Japanese and Hyun et al. (2020) for Korean.

Compared to English, to the best of our knowledge, no ABSA dataset for Czech could be used for compound ABSA tasks, e.g. the TASD task. Steinberger et al. (2014) introduced the first Czech ABSA dataset based on data from restaurant reviews, with the same type of annotations as in (Pontiki et al., 2014). Tamchyna et al. (2015) provide a dataset containing reviews of IT products with aspect term and sentiment polarity annotations. Unlike the Czech restaurant dataset by Steinberger et al. (2014), the IT product reviews are annotated with overall sentiment and aspect terms but lack categorization and sentiment classification for these terms. Hercig et al. (2016) expanded the Czech restaurant review ABSA dataset. The annotation format of the Czech restaurant datasets is based on the SemEval-2014 dataset and lacks linking aspect terms and categories. Moreover, these datasets have fewer, less detailed categories. For example, there is only a simple *food* category, and the *DRINKS* category is absent. However, the SemEval-2015 and SemEval-2016 datasets use categories in *E#A* format, combining entities and attributes, e.g. *FOOD#QUALITY* or *FOOD#PRICES*. See Figure 1 for an example.

2.2. Aspect-Based Sentiment Analysis

In recent years, there has been relatively little research on the ABSA task in Czech, and the existing approaches are often outdated compared to more modern sentiment classification methods.

The pioneering work on Czech ABSA was done by Steinberger et al. (2014), who presented baseline results for their restaurant dataset using Conditional Random Fields (CRF) and Maximum Entropy (ME) classifiers. Tamchyna et al. (2015) provided baseline results using CRF on their IT products dataset. Furthermore, Hercig et al. (2016) proposed various unsupervised methods to enhance performance in ABSA tasks for Czech and English, utilizing CRF and ME classifiers. Their research demonstrated that unsupervised methods could yield significant performance improvements. Lenc and Hercig (2016) employed a convolutional neural network (CNN) and recurrent neural network (RNN) for the task of identifying the sentiment polarity of aspect categories, evaluating their proposed model on the dataset from Hercig et al. (2016).

Most recently, Šmíd and Přibáň (2023) introduced the first prompt-based approach for SC and ABSA in Czech using models based on the Transformer architecture (Vaswani et al., 2017). One of their methods can solve multiple ABSA tasks simultaneously using sequence-to-sequence models. They also show the effectiveness of prompting in few-shot settings and that in-domain pretraining improves the results. Přibáň and Pražák (2023) combined ABSA tasks with the semantic information obtained by solving the semantic role labelling task. The multitask combination effectively improved results for Czech and English ACD and category polarity tasks.

To address the relative lack of work for Czech ABSA and provide an overview of the latest approaches, we present example studies focusing on ABSA in English, the most studied language in this field. Li et al. (2019) demonstrated the effectiveness of a BERT-based architecture for the E2E-ABSA task. Recent ABSA research primarily treats it as a text generation problem. Zhang et al. (2021b) introduced two paradigms for ABSA tasks designed to produce text output in a desired format, employing the English T5 model (Raffel et al., 2020). They achieved new state-of-the-art (SotA) benchmarks across various ABSA tasks, including TASD, using datasets from the SemEval competitions (Pontiki et al., 2014, 2015, 2016). Similarly, Zhang et al. (2021a) utilized the same English T5 model to address a recently introduced ABSA task known as aspect sentiment quad prediction, generating text output as well. Gou et al. (2023) introduced a method combining outputs with different orders of sentiment elements, showing that the order of elements matters, and achieved new SotA results on different datasets.

3. Dataset Construction

This paper aims to build a Czech dataset for ABSA within the restaurant domain, utilizing the annotation format consistent with SemEval-2015 (Pontiki et al., 2015) and SemEval-2016 (Pontiki et al., 2016) datasets. This annotation format is instrumental in addressing more complex ABSA tasks, including APD, ACTE and TASD. Additionally, aligning the annotation format with SemEval-2016 allows future cross-lingual experiments between Czech, English, and other languages in the SemEval-2016 dataset.

Our newly created dataset consists of two parts: a) manually annotated 3,189 reviews, see Tables 3 and 5 for detailed statistics and b) 24M raw additional reviews (330 MB of plain text) intended for additional unsupervised learning.

3.1. Unsupervised Dataset

The unsupervised part of the dataset was crawled from restaurant reviews on Google Maps[2]. Firstly, we obtained the list of names of Czech restaurants from the Restu.cz[3] website. Consequently, we searched each of the obtained restaurant names with Google Maps and downloaded all available reviews for the particular restaurant during September 2022. To maintain a certain level of anonymity, we provide only the reviews in the dataset. Additional details like the restaurant name, review date, or author's name are not included.

3.2. Manual Annotation

The manually annotated part of the dataset comprises two data segments. Firstly, we reuse all the reviews from the original dataset from Hercig et al. (2016). Secondly, we randomly sampled 1,110 reviews from the unsupervised part of the dataset. The existing annotations in the SemEval-2014 format cannot be directly converted into the SemEval-2016 format, and all the reviews must be read and labelled again from scratch. Thus, we started by completely reannotating the dataset from Hercig et al. (2016) following the annotations guidelines provided by Pontiki et al. (2015). Two native speakers annotated all the original data. For a given review, each annotator had the following tasks:

1. **Identify objective reviews:** Objective reviews and reviews without any sentiment expressed had to be marked[4] as *"OutOfScope"*. Example:

[2] http://googlemaps.com/
[3] https://restu.cz/
[4] To allow potential comparison and to keep the backward compatibility and consistency with the original dataset, we did not exclude the objective reviews.

"Koupila jsem 3 vouchery na pizzu." ("I bought 3 pizza vouchers.").

2. **Identify aspect terms:** One or more word expressions naming a specific aspect of the target entity, e.g. *"toast s vejci" ("toast with eggs")*. Implicitly referred aspect terms, e.g. in a review *"Doporučuji" ("Recommended")*, had to be assigned the value "NULL".
3. **Assign aspect category:** The annotators had to assign aspect categories for each identified aspect term. The aspect category consists of entity and attribute (*E#A*) and must be chosen from a pre-defined set of categories (e.g. *RESTAURANT#GENERAL* or *FOOD#QUALITY*). One aspect term could be assigned more aspect categories (for example, if the review mentions the quality and cost of the same aspect term). Example: *"Rychlá obsluha" ("Fast service") – "obsluha" ("service") → SERVICE#GENERAL*.
4. **Assign sentiment polarity:** Finally, for each (aspect term, aspect category) tuple, the annotators had to assign the sentiment polarity from one of the following values: *neutral*, *negative*, and *positive*. The *neutral* polarity applies to mildly negative or mildly positive sentiment.

An example of a review with annotation triplets compared to the annotations format used by Hercig et al. (2016) is shown in Figure 1.

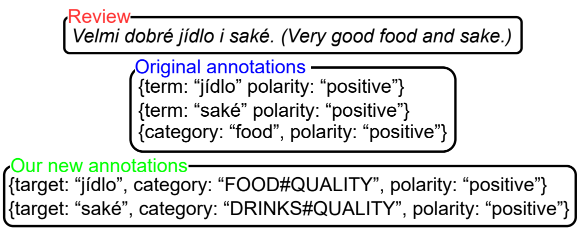

Figure 1: Our new annotations for a sample review compared to annotations from Hercig et al. (2016).

3.3. Annotators Details

Before the annotation of our dataset, all annotators have thoroughly passed the guidelines for annotations for SemEval 2015 (Pontiki et al., 2015) and SemEval 2016 (Pontiki et al., 2016) datasets and made a shared document with the important points regarding the annotation. Additionally, all annotators went through a few hundredths of annotated reviews from the English restaurant dataset from SemEval 2016 and made additional comments to the shared document. Then, all annotators discussed the main points.

Subsequent to the initial discussion, the annotators started the annotation process. During the annotation, after every few hundred new annotated data, the annotators reviewed the problems (if any) and went through the comments again. This procedure ensured the best possible annotator agreement and mitigated a lot of potential issues that might have occurred. Therefore, we only encountered those mentioned in Section 3.4.

3.4. Inter-annotator Agreement

Following Pontiki et al. (2016); Steinberger et al. (2014); Hercig et al. (2016), we calculated the inter-annotator agreement (IAA) as F1-score, where annotations from one annotator are treated as gold data and annotation from the second annotator as predictions. Table 2 shows the agreement. Similarly, Pontiki et al. (2016) achieved comparable results with values between 80 and 91% of IAA for the Spanish dataset. This fact indicates a similar level of quality of our dataset.

Annotation target	IAA
Aspect term	93.19
Aspect category	93.00
Aspect term & aspect category	91.06
Aspect term & aspect category & polarity	89.70

Table 2: Inter-annotator agreement (IAA) for different annotation targets in the new Czech dataset, measured in terms of micro F1 score (in %).

The main reasons for disagreements were mainly in the sentiment polarity, where, in some cases, it is difficult to determine whether the polarity is slightly positive or negative, hence neutral, or whether it should be assigned as strongly negative or positive. The definition of *RESTAURANT#GENERAL* and *RESTAURANT#MISCELLANEOUS* categories in the guidelines and datasets provided by Pontiki et al. (2015, 2016) are ambiguously defined. These two categories were another primary source of disagreement. The additional third annotator resolved the disagreement cases.

Following the approach described above, we additionally annotated 1,110 new examples of the randomly sampled reviews from the unsupervised part of the dataset. We then removed the randomly sampled reviews from the unsupervised dataset to avoid their potential use for training, as the newly annotated reviews may also be present in test data. We also removed all the original data we found in the unsupervised dataset.

Given that we considered the agreements substantially high for all annotation targets, we split all the new reviews not presented in the original datasets from Hercig et al. (2016) into two parts. Each annotator then independently annotated one part. Following annotation, each aspect's starting and ending offsets were automatically generated and labelled as "from" and "to" in the dataset.

We filtered out reviews without opinion triplets (i.e. objective reviews) and reviews in languages other than Czech from the new part. Example of filtered objective review: *"Bar navštěvovaný mladými lidmi" ("Bar frequented by young people")*. After this filtration, 1,040 reviews remained in the dataset.

3.5. Dataset Details

To enhance future research, we provide three versions of our dataset, named `CsRest`[5], with train, validation and test splits. The first version (`CsRest-O`) exclusively comprises the reannotated data from (Hercig et al., 2016), with 25% of this data designated as the test set. The other versions (`CsRest-N` and `CsRest-M`) contain all data, including newly annotated data not present in the original dataset (Hercig et al., 2016). In `CsRest-N`, all the new data serves as the test data, while in `CsRest-M`, we joined all the data together, shuffled them, and randomly selected 25% as the test data. For all three versions of the dataset, we further split the data not included in the test set into training and validation sets in a 9:1 ratio. The selection of 25% for test size is based on a similar value used in Pontiki et al. (2016).

We significantly expanded the original dataset by almost 50%, increasing the number of reviews from 2,149 to 3,189 for our dataset's `CsRest-N` and `CsRest-M` versions. This expansion introduced more than a 75% growth in the number of triplets, from 3,670 to 6,478, compared to the `CsRest-O` version of our dataset, which exclusively contains data from the original dataset. Compared to the SemEval-2016 datasets, the two larger versions (`CsRest-N` and `CsRest-M`) of the Czech ABSA dataset now stand as the second largest restaurant domain datasets regarding the number of reviews (only behind the Russian version) and the largest in the number of annotation triplets.

Table 3 shows the statistics of the dataset in terms of the number of reviews, annotation triplets, reviews without any annotation triplets and the number of "NULL" aspect terms (i.e. implicitly mentioned). Table 5 shows detailed statistics regarding aspect categories and sentiment polarity. We can see the imbalance in both aspect categories and sentiment polarity. The *neutral* sentiment polarity is the least frequent. The reviews most often mention the food quality and the restaurant and service. On the other hand, they do not often mention the location or prices. Table 4 shows a comparison between our dataset and those in other languages within the restaurant domain in terms of the number of reviews and annotated triplets.

[5]The versions suffix names come from O – **O**riginal, N – **N**ew and M – **M**ixed.

Split	Count	CsRest-O	CsRest-N	CsRest-M
Train	Reviews	1,450	1,934	2,151
	Triplets	2,510	3,240	4,386
	No triplets	104	142	109
	NULL terms	590	795	961
Dev	Reviews	162	215	240
	Triplets	253	430	483
	No triplets	6	17	9
	NULL terms	64	95	104
Test	Reviews	537	1,040	798
	Triplets	907	2,808	1,609
	No triplets	49	0	41
	NULL terms	49	517	342
Total	Reviews	2,149	3,189	3,189
	Triplets	3,670	6,478	6,478
	No triplets	159	159	159
	NULL terms	890	1,407	1,407

Table 3: Statistics of our dataset.

Dataset	Reviews	Triplets
en	2,676	3,366
es	2,951	3,792
fr	2,429	5,322
nl	2,286	2,473
ru	4,699	5,322
tr	1,248	1,694
CsRest-O	2,149	3,670
CsRest-N	3,189	6,478
CsRest-M	3,189	6,478

Table 4: Statistics of our dataset compared to datasets in another languages in the restaurant domain provided by Pontiki et al. (2016).

Our newly annotated dataset offers several improvements compared to the original dataset (Hercig et al., 2016). It links information between aspect terms and categories and aligns with the SemEval-2016 dataset, allowing us to perform more advanced tasks. It also provides more detailed annotations. For example, our annotations comprise entities and attributes in *E#A* format, e.g. *FOOD#QUALITY* or *FOOD#PRICES*, whereas the original dataset would merge them into a *food* category. Additionally, our dataset introduces new categories (entities) not present in the original dataset, such as "DRINKS".

4. Experiments & Setup

To evaluate the quality of the proposed dataset, we conduct experiments on the following tasks:

- Aspect term extraction (ATE): Extraction of aspect terms.

	Category	Train				Dev				Test				Total			
		Pos	Neg	Neu	Tot	Pos	Neg	Neu	Tot	Pos	Neg	Neu	Tot	Pos	Neg	Neu	Tot
CsRest-O	AMBIENCE#GENERAL	89	75	8	172	5	14	1	20	41	20	4	65	135	109	13	257
	DRINKS#PRICES	2	10	8	20	0	0	0	0	0	1	0	1	2	11	8	21
	DRINKS#QUALITY	61	30	16	107	4	6	1	11	22	14	2	38	87	50	19	156
	DRINKS#STYLE_OPTIONS	7	13	6	26	0	2	1	3	1	0	0	1	8	15	7	30
	FOOD#PRICES	18	38	11	67	0	3	2	5	2	16	3	21	20	57	16	93
	FOOD#QUALITY	400	275	87	762	32	29	12	73	166	113	16	295	598	417	115	1,130
	FOOD#STYLE_OPTIONS	49	84	41	174	5	4	2	11	10	33	7	50	64	121	50	235
	LOCATION#GENERAL	5	4	0	9	2	0	0	2	2	0	1	3	9	4	1	14
	RESTAURANT#GENERAL	278	229	35	542	23	31	2	56	99	99	13	211	400	359	50	809
	RESTAURANT#MISCELLANEOUS	5	52	25	82	3	6	6	15	2	24	1	27	10	82	32	124
	RESTAURANT#PRICES	21	27	9	57	1	4	1	6	13	14	4	31	35	45	14	94
	SERVICE#GENERAL	209	231	52	492	12	30	9	51	62	92	10	164	283	353	71	707
	Total	1,144	1,068	298	2,510	87	129	37	253	420	426	61	907	1,651	1,623	396	3,670
CsRest-N	AMBIENCE#GENERAL	112	99	11	222	23	10	2	35	316	25	22	363	451	134	35	620
	DRINKS#PRICES	0	8	6	14	2	3	2	7	10	13	2	25	12	24	10	46
	DRINKS#QUALITY	78	48	17	143	9	2	2	13	179	20	8	207	266	70	27	363
	DRINKS#STYLE_OPTIONS	8	14	6	28	0	1	1	2	42	7	5	54	50	22	12	84
	FOOD#PRICES	19	50	13	82	1	7	3	11	36	26	15	77	56	83	31	170
	FOOD#QUALITY	527	373	103	1,003	71	44	12	127	698	94	48	840	1,296	511	163	1,970
	FOOD#STYLE_OPTIONS	54	112	42	208	10	9	8	27	108	27	8	143	172	148	58	378
	LOCATION#GENERAL	5	3	0	8	4	1	1	6	36	3	0	39	45	7	1	53
	RESTAURANT#GENERAL	346	321	47	714	54	38	3	95	360	44	16	420	760	403	66	1,229
	RESTAURANT#MISCELLANEOUS	8	77	28	113	2	5	4	11	22	13	4	39	32	95	36	163
	RESTAURANT#PRICES	32	40	13	85	3	5	1	9	51	20	21	92	86	65	35	186
	SERVICE#GENERAL	248	310	62	620	35	43	9	87	404	81	24	509	687	434	95	1,216
	Total	1,437	1,455	348	3,240	214	168	48	430	2,262	373	173	2,808	3,913	1,996	569	6,478
CsRest-M	AMBIENCE#GENERAL	306	90	26	422	35	7	3	45	110	37	6	153	451	134	35	620
	DRINKS#PRICES	8	13	5	26	1	1	4	6	3	10	1	14	12	24	10	46
	DRINKS#QUALITY	181	51	15	247	20	5	2	27	65	14	10	89	266	70	27	363
	DRINKS#STYLE_OPTIONS	32	19	12	63	2	2	0	4	16	1	0	17	50	22	12	84
	FOOD#PRICES	38	55	18	111	7	6	2	15	11	22	11	44	56	83	31	170
	FOOD#QUALITY	882	366	116	1,364	86	39	9	134	328	106	38	472	1,296	511	163	1,970
	FOOD#STYLE_OPTIONS	119	91	41	251	15	10	6	31	38	47	11	96	172	148	58	378
	LOCATION#GENERAL	28	1	1	30	1	2	0	3	16	4	0	20	45	7	1	53
	RESTAURANT#GENERAL	525	264	44	833	55	38	5	98	180	101	17	298	760	403	66	1,229
	RESTAURANT#MISCELLANEOUS	23	64	22	109	2	7	2	11	7	24	12	43	32	95	36	163
	RESTAURANT#PRICES	58	47	24	129	3	4	2	9	25	14	9	48	86	65	35	186
	SERVICE#GENERAL	463	277	61	801	51	40	9	100	173	117	25	315	687	434	95	1,216
	Total	2,663	1,338	385	4,386	278	161	44	483	972	497	140	1,609	3,913	1,996	569	6,478

Table 5: Detailed statistics of our dataset regarding aspect categories and sentiment polarity.

- Aspect category detection (ACD): Detection of aspect categories.
- Aspect-category-term extraction: Extraction of (aspect term, aspect category) tuples.
- Aspect polarity detection (APD): Detection of sentiment polarity for given (aspect term, aspect category) tuples.
- End-to-end ABSA (E2E-ABSA): Extraction of (aspect term, sentiment polarity) tuples.
- Target-aspect-sentiment detection (TASD): Detection of (aspect term, aspect category, sentiment polarity) triples.

For all tasks, we use the micro F1-score as evaluation metrics, and following related work (Zhang et al., 2021a; Gou et al., 2023; Zhang et al., 2021b), we discard all examples without any annotations (i.e. objective reviews).

4.1. Encoder-Based Models

We use encoder-based (BERT-like) models to perform ATE, ACD, E2E-ABSA and APD tasks. We employ four Czech pre-trained Transformer-based models, specifically Czert (Sido et al., 2021), RobeCzech (Straka et al., 2021), FERNET (Lehečka and Švec, 2021) and Small-E-Czech (Kocián et al., 2022). Additionally, we use three multilingual models, specifically the multilingual BERT (mBERT) (Devlin et al., 2019) and the base and large version of XLM-RoBERTa (XLM-R) (Conneau et al., 2020). The encoder-based models convert an input sequence $x = w_1, \ldots, w_n$ of n tokens into a sequence of hidden vectors $\mathbf{h} = \mathbf{h}_0, \mathbf{h}_1, \ldots, \mathbf{h}_n$. The hidden vector $\mathbf{h}_0 = \mathbf{h}_{\texttt{[CLS]}}$ is the artificial classification `[CLS]` token representing the entire input sequence. For each task, we utilize a linear layer on top of the model to generate predictions and fine-tune the model's parameters $\mathbf{\Theta}$ that include task-specific parameters $\mathbf{W}$ and $\mathbf{b}$.

For the APD task, we create one input for each (aspect term, aspect category) tuple, where we append the tuple after the original review (the only task we solve where the number of inputs can be larger than the number of reviews). The linear layer computes the probability of a label y from a label space $\mathcal{Y} \in \{\textit{positive}, \textit{negative}, \textit{neutral}\}$ for the input x_i as

$$P_{\mathbf{\Theta}}(y|x_i) = \text{softmax}(\mathbf{W}\mathbf{h}_{\texttt{[CLS]}} + \mathbf{b}). \quad (1)$$

We choose the class with the largest probability.

The ACD task is similar to the APD task, but the label space is different; it contains all possible aspect categories. This task is also multi-label and not multi-class classification; hence, 0 to k classes can be selected instead of precisely one, where k is the total number of classes. We select all classes with a probability larger than 0.5.

To each token, a label is assigned for the ATE and E2E-ABSA tasks using BIO tagging, which denotes the aspect boundaries. For the ATE task, the class y_i for each token x_i comes from a label space $\mathcal{Y} \in \{B, I, O\}$, and for the E2E-ABSA task, from a label space $\mathcal{Y} \in \{B, I\} - POS, NEG, NEU \cup \{O\}$. For example, $y_i = B\text{-}NEG$ means x_i is the beginning of a negative aspect term. The label distribution of x_i is computed as

$$P_{\mathbf{\Theta}}(y_i|x_i) = \text{softmax}(\mathbf{W}\mathbf{h}_i + \mathbf{b}). \quad (2)$$

In the case of the E2E-ABSA task, if the same aspect term appears with different polarities in one review, we assign it the *neutral* polarity. Prediction for both tasks is considered correct only if the boundary (and sentiment polarity in the case of E2E-ABSA) are correct.

Three Czech models (Czert, RobeCzech and FERNET) are further pre-trained using the masked language modelling (MLM) task (Devlin et al., 2019) with the intention to adapt them to the task domain and improve the overall results.

4.2. Sequence-to-Sequence Models

We employ sequence-to-sequence models to perform ACD, ATE, ACTE and TASD tasks simultaneously. These models process text as input and produce text as output. To the best of our knowledge, no monolingual sequence-to-sequence models are currently explicitly designed for Czech. Consequently, we have decided to use the large **mT5** model (Xue et al., 2021) and the large **mBART** model (Tang et al., 2021). These models are the multilingual adaptations of the English T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) models.

The sequence-to-sequence models consist of two parts of the Transformer architecture: the *encoder* and the *decoder*. The encoder transforms input sequence x into a contextualized sequence $\mathbf{e}$. Given the encoded input $\mathbf{e}$, the decoder models the conditional probability distribution $P_{\mathbf{\Theta}}(y|\mathbf{e})$ of the target sequence y, where $\mathbf{\Theta}$ are the model's parameters. The decoder calculates the output y_i at each step i based on the previous outputs $y_1, \ldots, y_{i-1}$ and the encoded input $\mathbf{e}$.

Since the output of sequence-to-sequence models is text, we convert discrete ABSA labels to textual format, following Šmíd and Přibáň (2023). The label is constructed as "c *is* $P_p(p)$*, given the expression:* a", where a is the aspect term, c is the aspect category, and $P_p(p)$ is a mapping function that maps the sentiment polarity p as

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

For example, the review *"Výtečné pivo" ("Excellent beer")* yields the label "*Drinks quality* is *great*, given the expression: *pivo*". Multiple sentiment triplets in reviews are concatenated with semicolons.

In this context, the model takes the text (review) as input and aims to generate a textual label as its output. The model's parameters are fine-tuned to optimize label generation in the specified format. The model always generates all outputs together, i.e. for the TASD task, from which specific elements required for other tasks are extracted, e.g. category for the ACD task. We discard "NULL" targets for the ATE task and ignore duplicate targets for the ATE and ACD tasks, as in Pontiki et al. (2016).

Since our approach predicts the aspect category and term alongside sentiment, we do not use these models for the APD task, which assumes the model has access to the gold data for aspect terms and categories. A fair comparison would require modifications of the input and output formats. Similarly, we refrain from using these models for the E2E-ABSA task as their results cannot be fairly compared with encoder-based models. Sequence-to-sequence models can predict "NULL" terms and generate one aspect term multiple times with different polarities. In contrast, encoder-based models predict only one polarity for a single aspect term and do not predict the "NULL" aspect term.

4.3. Hyperparameters

We train the models with various hyperparameters. We use a batch size of 64 for each experiment and search for the learning rate from {3e-4, 1e-4, 5e-5, 1e-5}. Encoder-based models run for up to 50 epochs, while sequence-to-sequence models run for up to 35 epochs, using the greedy encoding algorithm for simplicity. We employ the AdamW optimizer (Loshchilov and Hutter, 2017) for all the models except the mT5 model, where we use the Adafactor optimizer (Shazeer and Stern, 2018).

We then select the best-performing models on validation data, fine-tune them on merged training and validation data and evaluate them on the test data. We conduct each experiment five times, each with a different random seed, to ensure the reliability of our results. We present the average scores along with a 95% confidence interval.

We also use the AdamW optimizer and the cross-entropy loss function for the additional MLM pre-training. The word masking probability is set to 15%. We pre-train the model for 20K steps with a batch size of 512 and a maximum input length set to 512 tokens with a learning rate of 5e-5.

5. Results

Table 6 shows the results achieved by the encoder-based models. We can see that the multilingual XLM-R models (particularly the large version, which has the most parameters out of all these models) perform similarly to Czech-only (monolingual) models. In some cases, they outperform them. The easiest task is the APD task, where the models assign only one of three classes. The ACD task is more complex than the APD task because the models have to choose from more classes, and the problem is multi-label. E2E-ABSA is the most challenging task because the model has to assign the correct class to each token and correctly predict the aspect term boundaries alongside the sentiment polarity. The ATE task is less difficult than the E2E-ABSA task because the model does not have to assign the sentiment polarity for the tokens. These claims are supported by the reported results corresponding to the different difficulty levels of each task; the easiest tasks achieve much better results than the more difficult ones. The baseline results shown in Table 6 achieved by Hercig et al. (2016) are on the old dataset, which has different annotations and aspect categories.

Overall, the results for the `CsRest-O` dataset are generally worse than for the two remaining datasets, possibly due to the smaller training data size. While there are some differences between the results for `CsRest-N` and `CsRest-M` datasets for each task and model, it is unclear whether either version is consistently more challenging.

Additionally, we pre-trained three Czech models on the unsupervised dataset. The results show that the additional pre-training significantly improves the performance of all three models. For example, the RobeCzech model shows an improvement of approximately 4% on the E2E-ABSA task and `CsRest-M` dataset.

Table 7 shows the results of sequence-to-sequence models. The mBART model outperforms the mT5 model on all tasks. The mT5 model also performs the best on the `CsRest-O` dataset compared to the other versions. The mBART model performs similarly on all versions. Worse results of the mT5 model could imply that a better hyperparameters search should be done for the mT5 model. Overall, the TASD task is the most challenging because the model must simultaneously predict the aspect term, aspect category and sentiment polarity correctly.

The encoder-based models consistently outperform the sequence-to-sequence models. The reason may be that the encoder-based models are always specialized directly for one task. On the other hand, the sequence-to-sequence models generate the output for the TASD task. Then, we extract only the relevant elements for the specific task from the output (e.g. only aspect terms for the ATE task).

5.1. Error Analysis

We conducted an error analysis of the sequence-to-sequence models to understand the key characteristics of our dataset and identify the main challenges these models face. Our findings revealed several important observations:

Output format: The mT5 model occasionally struggles to produce the correct output format, which is crucial for target extraction. On the other hand, the mBART model makes this error to a lesser extent, possibly contributing to its superior performance over mT5. Both models frequently generate duplicate outputs, reducing the diversity of generated sentiment triplets. While we ensure not to count identical triplets multiple times (thus not impacting the results), this repetition restricts the models from generating unique outputs, potentially causing them to miss specific prediction targets.

Aspect term prediction: Both models encounter challenges in predicting the correct aspect terms. They sometimes generate only a part of the aspect term rather than the complete term (e.g. *"burrito"* instead of *"burrito bowl"*). Additionally, the models may blend parts of the review, leading to outputs that do not match the original text's specific form. For example, instead of *"raspberries with ice cream and whipped cream"*, the model generates *"raspberries with whipped cream"*, a phrase not present in the original review.

Handling typos: The models generate words in the correct form even when they appear as typos in the original review. For instance, if the review contains the typo *"se**vr**ice"*, the model generates the corrected word *"service"*. The models also occasionally produce lowercase output even when the original text contains uppercase letters.

Making up words: The models sometimes make up words not found in the reviews. For example, some reviews imply opinions about the ambience, and the models may generate *"ambience"* instead of "NULL" as an aspect term.

Aspect category confusion: Regarding aspect categories, the models frequently omit the less common categories, such as *LOCATION#GENERAL* or *DRINKS#STYLE_OPTIONS*. Both models often confuse the *RESTAURANT#MISCELLANEOUS* and *RESTAURANT#GENERAL* classes.

Sentiment polarity challenges: The most significant challenge arises with neutral sentiment polarity. Despite being the least frequent class, both models rarely predict it and tend to predict either negative or positive sentiment.

Model	CsRest-O				CsRest-N				CsRest-M			
	APD	ACD	ATE	E2E	APD	ACD	ATE	E2E	APD	ACD	ATE	E2E
Czert	83.2$^{\pm1.4}$	81.2$^{\pm1.4}$	81.7$^{\pm0.4}$	66.8$^{\pm0.7}$	85.5$^{\pm4.9}$	81.6$^{\pm1.5}$	78.4$^{\pm1.0}$	70.9$^{\pm1.2}$	85.3$^{\pm0.9}$	82.2$^{\pm0.3}$	82.8$^{\pm0.7}$	70.6$^{\pm0.9}$
RobeCzech	85.2$^{\pm1.6}$	80.9$^{\pm2.5}$	82.9$^{\pm0.4}$	67.8$^{\pm1.6}$	89.4$^{\pm1.2}$	80.8$^{\pm1.6}$	78.8$^{\pm1.1}$	71.9$^{\pm1.6}$	87.6$^{\pm1.3}$	83.1$^{\pm1.0}$	82.8$^{\pm0.5}$	71.3$^{\pm1.9}$
FERNET	<u>86.0</u>$^{\pm0.4}$	<u>83.7</u>$^{\pm1.2}$	**84.9**$^{\pm1.1}$	<u>71.7</u>$^{\pm2.1}$	<u>90.1</u>$^{\pm2.2}$	**82.9**$^{\pm0.9}$	**80.8**$^{\pm1.2}$	<u>74.7</u>$^{\pm1.5}$	**88.2**$^{\pm0.8}$	<u>84.3</u>$^{\pm0.4}$	<u>83.2</u>$^{\pm1.1}$	**74.8**$^{\pm1.1}$
Small-E-Czech	78.0$^{\pm3.7}$	75.5$^{\pm1.2}$	81.5$^{\pm0.9}$	59.2$^{\pm4.8}$	84.6$^{\pm1.6}$	76.7$^{\pm2.0}$	77.3$^{\pm2.4}$	64.0$^{\pm2.9}$	83.3$^{\pm0.8}$	79.8$^{\pm0.7}$	81.1$^{\pm0.7}$	66.7$^{\pm2.0}$
mBERT	77.1$^{\pm3.8}$	77.8$^{\pm2.1}$	79.6$^{\pm0.6}$	60.3$^{\pm2.8}$	85.1$^{\pm1.8}$	78.6$^{\pm1.3}$	76.2$^{\pm1.7}$	67.5$^{\pm2.0}$	82.2$^{\pm1.0}$	79.0$^{\pm0.5}$	80.0$^{\pm0.6}$	67.7$^{\pm1.6}$
XLM-R$_{\text{BASE}}$	80.7$^{\pm2.3}$	80.4$^{\pm1.4}$	82.4$^{\pm0.6}$	68.9$^{\pm3.6}$	88.5$^{\pm1.8}$	80.6$^{\pm1.7}$	78.6$^{\pm1.3}$	70.7$^{\pm2.1}$	85.1$^{\pm1.6}$	81.0$^{\pm1.1}$	82.0$^{\pm0.4}$	70.4$^{\pm0.7}$
XLM-R$_{\text{LARGE}}$	**87.2**$^{\pm1.5}$	**85.7**$^{\pm0.4}$	<u>84.0</u>$^{\pm0.8}$	**71.9**$^{\pm2.3}$	**91.4**$^{\pm0.9}$	<u>82.8</u>$^{\pm1.0}$	<u>80.2</u>$^{\pm1.1}$	**75.5**$^{\pm1.0}$	<u>87.9</u>$^{\pm0.8}$	**86.2**$^{\pm0.3}$	**83.5**$^{\pm1.2}$	<u>74.4</u>$^{\pm1.0}$
Czert*	88.4$^{\pm0.7}$	86.8$^{\pm0.9}$	85.7$^{\pm1.7}$	74.7$^{\pm1.4}$	89.2$^{\pm2.6}$	84.6$^{\pm0.4}$	81.3$^{\pm1.4}$	73.8$^{\pm1.2}$	88.3$^{\pm1.1}$	86.1$^{\pm0.5}$	84.4$^{\pm1.0}$	75.6$^{\pm0.5}$
RobeCzech*	88.4$^{\pm0.9}$	84.9$^{\pm0.7}$	85.3$^{\pm1.1}$	70.4$^{\pm2.3}$	91.1$^{\pm0.8}$	83.9$^{\pm0.7}$	82.3$^{\pm1.0}$	74.3$^{\pm0.7}$	88.4$^{\pm0.8}$	85.7$^{\pm0.9}$	84.9$^{\pm1.2}$	75.4$^{\pm1.1}$
FERNET*	85.0$^{\pm1.1}$	83.9$^{\pm0.7}$	84.0$^{\pm0.9}$	71.7$^{\pm1.0}$	91.0$^{\pm1.5}$	84.0$^{\pm1.5}$	82.3$^{\pm1.2}$	75.9$^{\pm0.8}$	90.0$^{\pm0.5}$	87.1$^{\pm0.4}$	85.6$^{\pm0.8}$	77.0$^{\pm0.4}$
baseline†	-	80.0	78.7	-	-	-	-	-	-	-	-	-

Table 6: F1 scores for the new Czech ABSA dataset. The best score for each task and dataset version is in **bold**; the second best is <u>underlined</u>. Models marked with * are additionally pre-trained on the unsupervised dataset and are not considered for the best results. The † symbol denotes results by Hercig et al. (2016) obtained for the old dataset with different annotations and aspect categories.

Model	Task	CsRest-O	CsRest-N	CsRest-M
mT5	ACD	75.4$^{\pm1.8}$	68.9$^{\pm1.1}$	70.8$^{\pm1.5}$
	ATE	66.5$^{\pm2.5}$	59.7$^{\pm1.5}$	66.9$^{\pm1.4}$
	ACTE	56.4$^{\pm1.0}$	45.0$^{\pm1.7}$	52.6$^{\pm1.8}$
	TASD	48.0$^{\pm1.0}$	41.1$^{\pm1.8}$	46.4$^{\pm1.5}$
mBART	ACD	**78.7**$^{\pm1.6}$	**79.3**$^{\pm0.4}$	**80.6**$^{\pm1.7}$
	ATE	**78.9**$^{\pm1.3}$	**76.0**$^{\pm1.5}$	**79.7**$^{\pm1.1}$
	ACTE	**67.2**$^{\pm1.4}$	**62.4**$^{\pm0.7}$	**67.3**$^{\pm1.2}$
	TASD	**57.5**$^{\pm1.7}$	**56.3**$^{\pm0.6}$	**59.3**$^{\pm1.4}$

Table 7: F1 scores for different models across tasks on the new Czech ABSA dataset. The best result for each task and dataset version is in **bold**.

6. Conclusion

In this paper, we present a novel manually annotated Czech dataset in the restaurant domain for aspect-based sentiment analysis. The dataset comes in three different versions and is the largest of its kind in the Czech language. Unlike the previous Czech ABSA datasets, this newly created dataset establishes connections between multiple sentiment elements, allowing for solving more complex ABSA tasks, such as the TASD task. Notably, our dataset adheres to the same format as the SemEval-2016 dataset, potentially enabling cross-lingual experiments in the future. Next, we provide large unlabelled corpora for unsupervised training.

We also provide strong baseline results for various ABSA tasks utilizing models based on the Transformer architecture. Our system is language and domain-independent, meaning it can easily be trained on data from other languages. Our research extends beyond the numerical outcomes, delving into an insightful error analysis that elucidates the unique challenges and limitations our dataset poses to these models.

In summary, our study not only provides a new ABSA dataset for the Czech language but also establishes a benchmark for Czech ABSA research. We anticipate that this resource will catalyze future research endeavours, advancing the field of sentiment analysis and fostering cross-lingual exploration within the ABSA domain.

Acknowledgements

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:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

7. Bibliographical References

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.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent Twitter sentiment classification. In *Proceedings of the 52nd*

Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 49–54, Baltimore, Maryland. Association for Computational Linguistics.

Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. 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 2509–2518, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. MvP: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397, Toronto, Canada. 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.

Dongmin Hyun, Junsu Cho, and Hwanjo Yu. 2020. Building large-scale English and Korean datasets for aspect-level sentiment analysis in automotive domain. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 961–966, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6280–6285, Hong Kong, China. Association for Computational Linguistics.

Matej Kocián, Jakub Náplava, Daniel Stancl, and Vladimír Kadlec. 2022. Siamese bert-based model for web search relevance ranking evaluated on a new czech dataset. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 12369–12377. AAAI Press.

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.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, Hong Kong, China. Association for Computational Linguistics.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Yuki Nakayama, Koji Murakami, Gautam Kumar, Sudha Bhingardive, and Ikuko Hardaway. 2022. A large-scale Japanese dataset for aspect-based sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7014–7021, Marseille, France. European Language Resources Association.

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 and Ondřej Pražák. 2023. Improving aspect-based sentiment with end-to-end semantic role labeling model. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 888–897, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost.

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.

Jakub Šmíd and Pavel Přibáň. 2023. Prompt-based approach for Czech sentiment analysis. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1110–1120, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

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, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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.

Feixiang Wang, Man Lan, and Wenting Wang. 2018. Towards a one-stop solution to both aspect extraction and sentiment analysis tasks with neural multi-task learning. In *2018 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

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.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*.