Czech Dataset for Cross-lingual Subjectivity Classification

Pavel Přibáň[1,2], Josef Steinberger[1,2]

University of West Bohemia, Faculty of Applied Sciences,
[1]NTIS – New Technologies for the Information Society,
[2]Department of Computer Science and Engineering,
Univerzitní 2732/8, 301 00 Pilsen, Czech Republic
{pribanp,jstein}@kiv.zcu.cz
https://nlp.kiv.zcu.cz

Abstract

In this paper, we introduce a new Czech subjectivity dataset of 10k manually annotated subjective and objective sentences from movie reviews and descriptions. Our prime motivation is to provide a reliable dataset that can be used with the existing English dataset as a benchmark to test the ability of pre-trained multilingual models to transfer knowledge between Czech and English and vice versa. Two annotators annotated the dataset reaching 0.83 of the Cohen's κ inter-annotator agreement. To the best of our knowledge, this is the first subjectivity dataset for the Czech language. We also created an additional dataset that consists of 200k automatically labeled sentences. Both datasets are freely available for research purposes. Furthermore, we fine-tune five pre-trained BERT-like models to set a monolingual baseline for the new dataset and we achieve 93.56% of accuracy. We fine-tune models on the existing English dataset for which we obtained results that are on par with the current state-of-the-art results. Finally, we perform zero-shot cross-lingual subjectivity classification between Czech and English to verify the usability of our dataset as the cross-lingual benchmark. We compare and discuss the cross-lingual and monolingual results and the ability of multilingual models to transfer knowledge between languages.

Keywords: subjectivity, dataset, Czech, cross-lingual, classification, transformers, benchmark

1. Introduction

Subjectivity classification (Wiebe et al., 1999) is one of the integral parts of sentiment analysis (opinion mining). Its basic purpose is to determine if a sentence or phrase is subjective or objective (Liu, 2012). It can be further used to improve other tasks such as polarity detection or information extraction (Wiebe et al., 1999; Pang and Lee, 2004). Nowadays, the subjectivity classification is often used as a benchmark test (Zhao et al., 2015; Reimers and Gurevych, 2019; Wang et al., 2021; Bragg et al., 2021) in transfer learning to test abilities and language understanding of pre-trained BERT-like language models based on the Transformer architecture (Vaswani et al., 2017).

The goal of subjectivity classification is to classify a sentence or a clause of the sentence as subjective or objective. Subjective sentences express personal feelings, views, beliefs or opinions and objective sentences hold or describe some factual information (Liu, 2012). Evaluation of the pre-trained models for transfer learning is a crucial part of their development. For English, the well-known GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks are available. These benchmarks contain a set of diverse tasks that allow a thorough evaluation of English pre-trained models.

For multilingual models such as mBERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020), the XTREME (Hu et al., 2020) benchmark can be used to test their ability of cross-lingual transfer learning and knowledge transfer between languages. Unfortunately, the XTREME benchmark does not include any task for the Czech language.

Our main motivation is to partly fill this gap and contribute a bit by introducing a reliable Czech dataset that can be used for the cross-lingual evaluation. We intend to use the dataset to test the cross-lingual abilities of pre-trained multilingual models in pair with the existing English dataset (Pang and Lee, 2004) as a benchmark for zero-shot cross-lingual subjectivity classification. Thus, partly test the ability of pre-trained multilingual models to transfer knowledge between Czech and English. We are aware that to properly evaluate any pre-trained model, a diverse set of tasks is needed, but we believe that even one task can be helpful in the evaluation process. To the best of our knowledge, there is no subjectivity dataset for the Czech language, therefore our secondary goal is to extend the available dataset resources for Czech.

In this paper, we present the first Czech dataset for subjectivity classification task that consists of 10k manually annotated sentences from movie reviews and movie descriptions. Secondly, we provide an additional dataset of 200k sentences labeled in a distant supervised way (automatically). The automatic labeling is based on the idea from (Pang and Lee, 2004) that movie reviews contain mostly subjective sentences and the movie descriptions usually consist of objective sentences. We describe the process of building and annotating the dataset. The dataset is annotated by two annotators and the Cohen's κ (Cohen, 1960) inter-annotator agreement between them reaches 0.83. We

perform experiments with two multilingual mBERT (Devlin et al., 2019) and XLM-R-Large (Conneau et al., 2020) and three monolingual Transformer based models on the new Czech dataset and providing a competitive baseline of 93.56% of accuracy. Next, we conduct experiments with the same two multilingual models on the English dataset to be able to compare our cross-lingual experiments. Our results for the monolingual experiments with English are on par with the current state-of-the-art results. Finally, we evaluate the multilingual models and their ability to transfer knowledge between English and Czech on the zero-shot cross-lingual classification task. The cross-lingual experiments show that using only English data for fine-tuning the XLM-R-Large, the model can achieve worse results only by 2.8% on the Czech dataset compared to the model trained on Czech data. When the model is trained using only the Czech data, the result on the English dataset is roughly 4.4% worse than the current state-of-the-art results.

Our main contributions are the following: 1) we introduce the first Czech subjectivity dataset that allows cross-lingual evaluation in pair with the existing English dataset. 2) We perform a series of monolingual and cross-lingual experiments. We set a competitive baseline for the new Czech dataset. We compare abilities of two multilingual models to transfer knowledge between Czech and English in the subjectivity classification task. 3) We release[1] the dataset and code freely for research purposes, including the dataset splits for easier comparison and reproducibility of our results.

2. Related Work

The subjectivity classification task was a popular research topic at the beginning of the 21st century. It was studied in many papers (Wiebe and Wilson, 2002; Wiebe et al., 2004; Riloff et al., 2005; Esuli and Sebastiani, 2006; Wiebe and Mihalcea, 2006; Mihalcea et al., 2007). Nowadays, the subjectivity classification is often used as a benchmark for the evaluation of pre-trained models intended for transfer learning.

In (Wiebe et al., 1999), the authors describe the annotation process of 1k news sentences. Four annotators annotated the sentences as subjective or objective, but because some sentences can be considered borderline examples, they also assigned certainty ratings, ranging from 0, for least certain, to 3, for most certain. We use special label trash for the borderline sentences during our annotation, see Section 3.2.1.

Pang and Lee (2004) created English subjectivity dataset from movie reviews and movie descriptions. They automatically made the dataset using the assumption that reviews have mostly subjective sentences and descriptions usually contain objective sentences. The resulted dataset consists of 10k sentences, see Table 2.

Further in this paper, we reference the dataset as the English dataset.

In terms of Czech resources, the Czech subjectivity lexicon *Czech SubLex 1.0* (Veselovská and Bojar, 2013) contains a list of words with assigned sentiment polarity and part-of-speech tags.

There are also pairs of existing datasets that can be used for the cross-lingual evaluation similarly to our approach. For example, the Czech sentiment dataset of movie reviews *CSFD* (Habernal et al., 2013) can be used with the English *IMDB* (Maas et al., 2011) sentiment reviews dataset as shown in (Přibáň and Steinberger, 2021). Another example is the multilingual corpus (Piskorski et al., 2019) for named entity recognition (NER) that contains labels in the same format for four Slavic languages, including Czech. Next, the Czech aspect-based sentiment dataset (Hercig et al., 2016) can be evaluated together with the English dataset (Pontiki et al., 2014) and both of them come from the same domain and contain the same set of labels.

The initial work focused on cross-lingual subjectivity classification is presented in (Mihalcea et al., 2007). The authors investigated methods to automatically generate resources for subjectivity analysis for new language by using a parallel corpus and available resources in English. Amini et al. (2019) performed cross-lingual subjectivity classification between English and Persian. Other work that is related to cross-lingual subjectivity can be found in (Saralegi et al., 2013).

In (Wang et al., 2021), the authors used the English subjectivity dataset as one of the tasks to evaluate their approach for few-shot learning based on RoBERTa model (Liu et al., 2019). (Nandi et al., 2021) analyzed various models for text representation, including the original BERT model (Devlin et al., 2019) on the English dataset. Similarly, in (Zhao et al., 2015; Amplayo et al., 2018; Khodak et al., 2018; Reimers and Gurevych, 2019), the authors also used the English dataset to evaluate the performance of their newly designed models.

3. Dataset Building

We provide two datasets[1] of subjective and objective Czech sentences from movie reviews and movie descriptions (plot summaries), respectively. We use the mentioned idea from Pang and Lee (2004), in which the authors automatically created English dataset (`Subj-EN`) of 10k subjective and objective sentences. They assume that the descriptions are mostly objective and the reviews are subjective. This assumption is valid in most cases, but there can also be objective sentences in reviews and subjective sentences in descriptions. The number of these noisy samples differs in both cases, as you can see in Table 1.

For this reason, we decided to create a manually annotated dataset (`Subj-CS`) of 10k examples that should
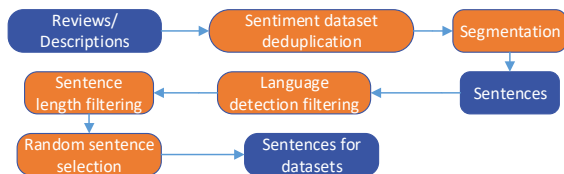
Figure 1: Data cleaning pipeline visualization.

eliminate the incorrect occurrences as much as possible. Secondly, we automatically built an additional dataset (Subj-CS-L) of 200k sentences using almost the same approach[2] as in (Pang and Lee, 2004).

3.1. Cleaning and Obtaining Data

We acquired roughly 4M reviews and 735k descriptions from Czech Movie Database[3] (CSFD) during October 2021. The Czech sentiment movie review dataset (Habernal et al., 2013) also consists of reviews from CSFD. We assume that in the future, our dataset can be used in combination with the sentiment dataset therefore, we decided to remove the sentiment reviews from the data downloaded by us. We were able to match and remove about 74k reviews out of a total of 91k from the sentiment dataset. The remaining 17k reviews were most likely changed or removed from the CSFD website since the authors of the sentiment dataset originally downloaded the data in 2013.

Next, we split the reviews and descriptions into sentences by UDPipe 2 (Straka, 2018)[4]. We have to note that in some cases, it failed to split the sentences correctly, especially for sentences without a space after the first sentence.

The reviews can contain phrases instead of grammatically correct sentences, but we do not distinguish between them. Some of the texts (mostly reviews) were written in other languages (most often Slovak and English). We filter these out[5] and we keep only Czech sentences. Finally, we filter out sentences with less than six tokens. See Figure 1 for the cleaning pipeline visualization.

The entire cleaning process resulted in 884k and 19M sentences (phrases) from descriptions and reviews, respectively. We randomly selected 40k sentences from the obtained reviews and descriptions for manual annotation and 200k sentences (100k from reviews and 100k from descriptions) for the automatically created dataset. The remaining sentences are not used.

---

[2]Based on our observations in the dataset, we decided to use sentences or phrases with at least six tokens but they used sentences longer than nine tokens.

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

[4]We use the *czech-pdt-ud-2.5-191206.udpipe* model.

[5]We use the Python package langdetect available at https://pypi.org/project/langdetect/ to detect the language.

3.2. Data Annotation

In this section, we describe the process of manual annotation. Two Czech native speakers performed the annotation. Even though the subjectivity classification may seem like an easy task, it showed to be rather difficult for some sentences to assign a subjectivity label.

3.2.1. Annotation Procedure

Firstly, the task of subjectivity classification was explained to the annotators along with the meaning of the subjective and objective sentences according to the definition in (Liu, 2012). The annotators were also asked to read the papers mentioned in Section 2 to clearly understand the task. We summarize the annotation guidelines in Section 3.2.3.

During the first annotation stage, each of the annotators was asked to label a common set of 100 sentences with one of three labels: subjective, objective and trash, see Section 3.2.3 for their description.

We use the trash label because, despite our best data cleaning efforts, there were still undesirable texts: e.g., short sequences of words that does not make any sense (random words), only numbers and other characters, sentences in other languages, texts that were obviously incorrectly segmented and made no sense etc.

After the first 100 annotated sentences, the annotators discussed the conflicts to clarify and improve the annotation guidelines. Based on the discussion, we decided to extend the annotation labels by two more unsure and question.

The questions appeared to be rather problematic. The subjectivity was not clear very often and thus, we decided to exclude them. In addition, the questions are only in a tiny part of the data, i.e., 1.73% and 2.41% for review and description sentences, respectively, see Table 1.

The unsure label was added because for some sentences, the annotators were not able to assign the subjectivity. For example, sentences for which a context (previous sentence) is needed to decide, sentences that describe a movie or event, but contain some clearly subjective adjective(s) and they can be perceived or interpreted both as subjective or objective depending on an individual person. Other problematic sentences are commands, wishes or parts of poems and rhymes. Here we list some of the problematic sentences that both annotators labeled with the unsure label:

(1) *"Všechno ovšem tak snadné řešení nemá."* – *"Not everything has such an easy solution."*

(2) *"To je dobrý důvod pro to, aby byla Japonsku vyhlášena válka."* – *"That's a good reason to declare war on Japan."*

(3) *"Dnes večer je to však díky napjaté atmosféře velmi obtížné."* – *"Tonight, the tense atmosphere makes it very difficult."*

(4) *"Drastický horor, při kterém tuhne krev v žilách"* – *"Drastic horror that makes your blood run cold"*

(5) *"Tak se o to postará příroda sama!"* – *"Nature will take care of it!"*

We decided to add these additional labels because we wanted to assign labels only in cases where the annotators are very confident with their annotations and thus obtain more reliable annotations without controversial examples and dataset of high quality.

After the update of the annotation guideline, both of the annotators assigned labels to the same 2,034 sentences. The Cohen's κ (Cohen, 1960) inter-annotator agreement for this 2k sentences reaches 0.68 for all five labels. Because we provide the dataset only with the objective and subjective labels, we exclude any sentence with at least one[6] of the trash, unsure or question labels. Thanks to this filtration, we obtain 1,668 sentences only with the subjective and objective labels. The Cohen's κ for this subset is 0.83, which represents a fairly good level of agreement. The remaining 141 conflict sentences are then resolved with the help of third person.

Finally, almost 5,000 sentences were annotated by each of the two annotators resulting in a total of 11,907 annotated sentences, see Table 1. We can see that the subjective and objective sentences are relatively balanced in the annotated samples and we believe that this reflects the real data distribution. Even though we obtained more than 5,000 sentences with the subjective and objective labels, we cut the annotations to have exactly 5,000 examples for each of the two labels. We decided to provide a perfectly balanced dataset since it allows easier comparison and evaluation of experiments. In our experiments, we use only the sentences with the subjective and objective labels, i.e., 10,000 sentences. We refer to this dataset as `Subj-CS`.

The entire procedure of annotation can be summarized into the following steps:

1. Each annotator annotated 100 sentences as `subjective`, `objective` or `trash`.

2. Every conflict in the first 100 sentences was discussed separately between the annotators to clarify and improve the annotation guideline. We extended the annotation guideline by two more labels: `unsure` and `question`.

3. 2,034 sentences are annotated by each annotator (1,668 as subjective or objective with 141 conflicts). The Cohen's κ reaches 0.83 for subjective and objective sentences. The conflicts are resolved by a third person.

4. Almost 10k another sentences are annotated in total by both annotators. The annotations are cut down to contain exactly 5,000 subjective and objective sentences.

---

[6]Each sentence has two labels – one from each annotator.

3.2.2. Annotation Statistics

The manual annotation resulted in a total of 11,907 annotated sentences, see Table 1. During the annotation procedure, we set the limit of at most 15 review sentences for the same movie and at most three description sentences in the 40k sentences selected for the manual annotation. However, the average number of sentences for the same movie is only 1.43 and 1.02 for review and description sentences, respectively.

Label	Reviews	Descriptions	Total
unsure	866 / 13.11%	457 / 8.62%	1 323
object.	**726 / 10.99%**	**4 464 / 84.22%**	**5 190**
subj.	**4 794 / 72.57%**	**208 / 3.92%**	**5 002**
quest.	114 / 1.73%	128 / 2.41%	242
trash	106 / 1.60%	44 / 0.83%	150
Total	6 606 / 100%	5 301 / 100%	11 907

Table 1: Annotation statistics for subjective and objective

As we assumed, a considerable percentage of sentences in reviews are not subjective (only 72.57% of sentences are subjective). Similarly, there is also a relatively large part of sentences in the movie descriptions that are not objective (84.22% of the sentences are objective).

3.2.3. Annotation Guideline

The annotators were instructed to annotate a given sentence or phrase with one of five labels. Based on the subjectivity description from (Wiebe et al., 1999; Pang and Lee, 2004; Liu, 2012), the sentence should be annotated as subjective if it expresses or evokes some personal feelings, views, beliefs or the sentence holds an opinion about entities, events or their properties (mostly movies in our case) from the non-objective point of view. For example:

"Samotný film se mi líbil, ale nepřekvapil." – *"I liked the movie itself, but it didn't surprise me."*

The sentence should be annotated as objective if it contains some factual information about an entity, event or their properties but does not hold a personal or subjective opinion about it and it does not try to convince or impose some opinion to the reader, for example:

"Maurice žije a pracuje v jižní Francii." – *"Maurice lives and works in the south of France."*

The disputed and controversial sentences, sentences where the annotator is not sure about its subjectivity or sentences for which context from previous text is needed to decide should be annotated with the unsure label, see Section 3.2.1 for examples. The trash label is used for sentences or phrases that do not make any sense or contain random words, characters or numbers. The question label is used for sentences that are questions.

3.3. Automatic Dataset

Besides the manually annotated dataset, we also built a large dataset (named `Subj-CS-L`) in a distant supervised way using the same approach as in (Pang and Lee, 2004). We labeled 100k review sentences as subjective and 100k movie description sentences as objective ones. All sentences have to have at least six tokens. We believe that even if the dataset contains some incorrect labels, it could be useful in combination with the manually created dataset, for example, in an unsupervised pre-training.

4. Data & Models for Experiments

For the experiments, we split the `Subj-CS` dataset into three parts with the following ratio: 75% for training, 5% for the development evaluation and 20% for testing. For the cross-lingual experiment with the `Subj-CS-L` dataset from Czech to English, we use 5% as the development evaluation data and the rest is used for training. Because there is no official split for the English dataset (Pang and Lee, 2004), we use 10-fold cross-validation for the monolingual experiments to be able to compare our results with other papers. We also split the English dataset into training, development and testing parts with the same test size (see Table 2) that was used in (Wang et al., 2021)[7]. For all three Czech and English datasets, we provide a script to obtain exactly the same data split to allow reproducibility and future comparison of our results.

Dataset	Name	Subjective	Objective	Total
Subj-CS	cs-train	3 750	3 750	7 500
	cs-dev	250	250	500
	cs-test	1 000	1 000	2 000
		5 000	5 000	10 000
Subj-CS-L	cs-L-train	95 000	95 000	190 000
	cs-L-dev	5 000	5 000	10 000
		100 000	100 000	200 000
Subj-EN	en-train	3 764	3 736	7 500
	en-dev	231	269	500
	en-test	1 005	995	2 000
		5 000	5 000	10 000

Table 2: Datasets statistics.

4.1. Transformer Models

For the experiments, we use solely the pre-trained BERT-like models based on the encoder part of the original Transformer architecture (Vaswani et al., 2017). The modified language modeling task is used to pre-train all the models, see the corresponding papers for details.

We employ three Czech monolingual models *Czert-B* (Sido et al., 2021), *RobeCzech* (Straka et al., 2021),

---

Czech Electra model (Kocián et al., 2021), two multilingual models *mBERT* (Devlin et al., 2019), *XLM-R* (Conneau et al., 2020) and the original monolingual English *BERT* model (Devlin et al., 2019), see Table 3 for their size (in a number of parameters) comparison.

Model	#Params	Vocab	#Langs
Czech Electra	13M	30k	1
Czert-B	110M	30k	1
RobeCzech	125M	52k	1
BERT	110M	30k	1
mBERT	177M	120k	104
XLM-R-Large	559M	250k	100

Table 3: A comparison of used models: number of parameters, vocabulary size and a number of supported languages.

Czech Electra (Kocián et al., 2021) is Czech model based on the Electra-small model (Clark et al., 2020).

Czert-B (Sido et al., 2021) is Czech variant of the original BERT$_{BASE}$ model (Devlin et al., 2019).

RobeCzech (Straka et al., 2021) is Czech version of the RoBERTa model (Liu et al., 2019).

BERT (Devlin et al., 2019) is the original BERT$_{BASE}$ model.

mBERT (Devlin et al., 2019) is a cased multilingual version of the BERT$_{BASE}$ that was jointly trained on 104 languages.

XLM-R-Large (Conneau et al., 2020) is a multilingual version of the RoBERTa (Liu et al., 2019) that supports 100 languages.

We fine-tune all the models for the binary classification task, i.e., subjective vs. objective sentence detection. For all models based on the original BERT model, we use the hidden vector $\mathbf{h} \in \mathbb{R}^H$ of the classification token `[CLS]` that represents the entire input sequence, where H is the hidden size of the model. The vector is obtained from the pooling layer, i.e., from a fully-connected layer of size H with a hyperbolic tangent used as the activation function. The dropout of 0.1 is applied and the result is then passed into a task-specific linear layer represented by matrix $\mathbf{W} \in \mathbb{R}^{|2| \times H}$. The output class c (subjective or objective) is computed as $c = \mathrm{argmax}\,(\mathbf{h}\mathbf{W}^T)$.

For the XLM-R-Large and RobeCzech models, the same[8] approach is used and in addition, an extra dropout of 0.1 is applied before the pooling layer (as in the original RoBERTa implementation).

We use the Adam (Kingma and Ba, 2014) optimizer with default parameters ($\beta_1 = 0.9, \beta_2 = 0.999$) and the cross-entropy loss function.

---

5. Experiments

To set baseline results for the new Czech dataset and verify its usability as a cross-lingual benchmark dataset between Czech and English, we performed a series of experiments with Transformer based models. The experiments can be categorized into two groups – *monolingual* and *cross-lingual*.

In monolingual experiments for Czech, we fine-tune the three Czech monolingual BERT-like models, i.e., *Czert-B*, *RobeCzech* and *Czech Electra* model and two multilingual models *mBERT* and *XLM-R*. For English, we use the same two multilingual models and the original *BERT* model. In *cross-lingual* experiments, we test the ability to transfer knowledge between Czech and English using the *zero-shot cross-lingual* classification. We fine-tune the multilingual models only on the dataset in one language (Czech or English) and then evaluate the fine-tuned model on the dataset in the other language.

We always fine-tune[9] on training data and measure the results on the development and testing data parts. We select the model that performs best on the development data and we report the results using average accuracy with the 95% confidence intervals (we repeat each experiment at least 12 times). We fine-tune all parameters of the model, including the added classification layers. We run the experiments for at most ten epochs with the linear learning rate decay (without learning rate warm-up) with the initial learning rates ranging from 2e-7 to 2e-4. The batch size is set to 32 and the max sequence length of the input is 200 since we classify sentences and the vast majority of them fit into this length. See Appendix A.1 for the hyper-parameters details for the reported experiment results.

5.1. Czech Monolingual Experiments

For Czech monolingual experiments, we use two types of training data. The training part (`cs-train`) of the manually labeled dataset `Subj-CS` and the entire automatically created dataset `Subj-CS-L` (marked as `cs-L-train`). In both cases, we evaluate models on the development (`cs-dev`) and testing (`cs-test`) parts of the `Subj-CS` dataset. We report the results in Table 4.

As we expected, the XLM-R-Large model achieves the highest average accuracy of 93.56% for both types of training data. Despite the highest achieved accuracy, there is an intersection in its confidence interval with RobeCzech model for the `cs-train` data (the * symbol in Table 4). Thus, we can conclude that RobeCzech and XLM-R-Large perform very similarly for Czech monolingual experiments. Thanks to the XLM-R-Large size (and its relatively large hardware training requirements), one could prefer the smaller

---

[9]The composition of data used for training and evaluation depends on the corresponding experiment. In the case of English monolingual experiments for the 10-fold split, we did not use any development data.

| Model | Subj-CS (cs-train) | Subj-CS-L (cs-L-train) |
	cs-test	cs-test
Czech Electra	91.85 ± 0.27	91.21 ± 0.08
Czert-B	92.85 ± 0.20	91.79 ± 0.07*
RobeCzech	93.29 ± 0.19*	91.63 ± 0.08
mBERT	91.23 ± 0.21	91.14 ± 0.11
XLM-R-Large	**93.56 ± 0.13**	**91.96 ± 0.10**

Table 4: Results for Czech monolingual experiments reported as average accuracy for the testing `cs-test` data part. The * symbol denotes results containing intersection in confidence interval with the best model.

RobeCzech model. The last observation is that all the models achieve better results with the `cs-train` data part. We expected XLM-R-Large to perform very well because it is the largest model and as shown in (Přibáň and Steinberger, 2021) it usually outperforms smaller monolingual models.

5.2. English Monolingual Experiments

In our English monolingual experiments, we evaluate the English dataset on our training (`en-train`), development (`en-dev`) and testing (`en-test`) data split. Because models from other works (Zhao et al., 2015; Amplayo et al., 2018; Khodak et al., 2018; Reimers and Gurevych, 2019; Nandi et al., 2021) are evaluated on the 10-fold split, we evaluate the models also on the 10-fold split (`en-10-fold`) to be able to compare their and ours results.

Model	en-test	en-10-fold
BERT	96.55 ± 0.16	96.87 ± 0.25
mBERT	95.87 ± 0.13	96.03 ± 0.24
XLM-R-Large	**97.28 ± 0.07**	**97.34 ± 0.21**
(Wang et al., 2021)†	97.40 ± 0.10*	-
(Nandi et al., 2021)	-	97.30
(Zhao et al., 2015)	-	95.50
(Amplayo et al., 2018)	-	94.80
(Khodak et al., 2018)	-	94.70
(Reimers and Gurevych, 2019)	-	94.52

Table 5: Results for English monolingual experiments reported as average accuracy for the testing `en-test` and `en-10-fold` data parts. The model in paper marked with the † symbol uses the same test size, but the distribution of sentences is different in each split part and they also use the standard deviation instead of the confidence interval.

As shown in Table 5, the XLM-R-Large performs best among the other two transformer models without any intersection of confidence intervals between the different models. We can also see that the results for `en-test` and `en-10-fold` are very similar and their confidence intervals overlap for the same model pairs (but different training data). Based on this observation, we assume that the results for `en-test` and `en-10-fold` are comparable to each other, thus in the cross-lingual experiments, English is evaluated

only on the `en-test` part.

We compare our results with the current state-of-the-art results (rows below the dashed line in Table 5). Most of the other works use the 10-fold cross-validation and our results also achieve the SotA results and are on par with them. We have to note that our 10-fold splits are not exactly the same as those in the referenced works because the authors do not provide them publicly. Using their distribution, we would probably get slightly different results. Nevertheless, we believe that we can compare our results with the other works to some extent.

5.3. Cross-lingual Experiments

We perform three types of cross-lingual experiments: from English to Czech, from Czech to English and joint training and evaluation of both languages. The first two are also known as a zero-shot cross-lingual classification because the model is fine-tuned only on data from one language (source language) and evaluated on data from the second language (target language). The model has never seen the labeled data from the target language.

For the experiments from English to Czech (EN→CS), we fine-tune the multilingual models on English `en-train` data and we evaluate them on the `en-dev` and `cs-test`. We select the model that performs best on the `en-dev` (i.e., the same best model as for the English monolingual data) and we report results for the `cs-test` data in Table 6[10].

| Model | EN → CS | | Monoling. (cs-train) |
	en-dev	cs-test	cs-test
mBERT	95.38 ± 0.22	86.18 ± 0.33	91.23 ± 0.21
XLM-R-Large	97.60 ± 0.18	**90.75 ± 0.32**	93.56 ± 0.13

Table 6: Accuracy results for cross-lingual experiments from English to Czech along with the results for models trained on monolingual data.

The XLM-R-Large model clearly outperforms the mBERT model by 4.5% but is worse than the same model that was trained on monolingual data roughly by 2.8%. In the case of mBERT, the results are much worse (5% difference) than the model trained only on monolingual data.

For experiments from Czech to English (CS→EN), we fine-tune the models on `cs-train` and evaluate on `cs-dev` and `en-test`. We select the model that performs best on `cs-dev`.

We also train the model on the `cs-L-train` data, but in this case, we select the model that performs best on the `en-dev` data from the target language (English). We use the `en-dev` for selecting the best model because we found out that if we use `cs-L-dev`, we get much worse results (up to 20% worse) for the

---

[10] We also include the monolingual results for an easier comparison of the results.

`en-test`. We are aware of this simplification of the zero-shot cross-lingual classification task, but otherwise, we would not be able to obtain a model with reasonable results. The results are stated in Table 7.

For both models trained on Czech data (`cs-train` and `cs-L-train`), the results are even worse in comparison to the previous experiment from English to Czech. For example, the difference between XLM-R-Large trained on `cs-train` and XLM-R-Large trained on English `en-train` data is 4.4%, whereas in the case of the previous experiment from English to Czech, it was only 2.8%. The results of the models trained on the `cs-L-train` are significantly worse (10% for mBERT).

Finally, we fine-tune the models jointly on `cs-train` and `en-train`, i.e., on both languages at once. We average the results obtained on `cs-dev` and `en-dev` and we select the model that achieves the highest average value. We report the results for the `cs-test` and `en-test` in Table 8.

We can see that the obtained results are almost identical or slightly different compared to the models trained only on monolingual data. Thus, we can conclude that the joint fine-tuning has no beneficial contribution.

5.4. Discussion

In this section, we summarize and mention some of our main findings and conclusions from the experiments. Even though that the Czech Electra model is significantly smaller than all the other models, it achieves very competitive results compared to the other models. Thanks to its smaller size, it is much easier and faster to be fine-tuned.

The XLM-R-Large model dominates the results, but it is also several times larger than the other models, see Table 3. Despite the worse results in the cross-lingual experiments, we can state that generally, the XLM-R-Large (and in some cases even mBERT) is relatively capable of transferring knowledge between Czech and English and vice versa, at least for the subjectivity classification task. The confidence intervals for results obtained in cross-lingual experiments are usually larger than the ones for the monolingual results. Thus, we consider the cross-lingual results less stable.

During the cross-lingual experiments, we select the best model based on development results for the source language. We believe that this is more difficult and challenging than choosing the model according to the results on the target language. We also believe that this setting is much closer to the potential usage of the multilingual models in the industry or to solving practical real-world tasks that are often more complicated. We do not use this approach for models trained on the large data that were obtained automatically because of its poor results.

Based on the cross-lingual results, we believe that for knowledge transfer between languages, a smaller but high-quality (manually annotated) dataset is better

| Model | CS → EN (cs-train) | | CS → EN (cs-L-train) | | Monolingual (en-train) |
	cs-dev	en-test	en-dev	en-test	en-test
mBERT	92.11 ± 0.38	88.99 ± 0.94	85.80 ± 0.89	85.53 ± 0.98	95.87 ± 0.13
XLM-R-Large	94.40 ± 0.36	**92.86 ± 0.44**	93.35 ± 0.22	**90.98 ± 0.26**	97.28 ± 0.07

Table 7: Accuracy results for cross-lingual experiments from Czech to English along with the results for models trained on monolingual data.

| Model | Joint (cs-train + en-train) | | Monolingual (cs-train) | Monolingual (en-train) |
	cs-test	en-test	cs-test	en-test
mBERT	91.12 ± 0.24	95.69 ± 0.22	91.23 ± 0.21	95.87 ± 0.13
XLM-R-Large	**93.85 ± 0.15**	**96.95 ± 0.12**	93.56 ± 0.13	97.28 ± 0.07

Table 8: Accuracy results for models jointly trained on English and Czech data along with the results for models trained on monolingual data.

and more important than a large automatically created dataset to obtain more reliable results for downstream tasks.

6. Conclusion

In this work, we introduce the first Czech subjectivity dataset `Subj-CS` that consists of 10k manually annotated subjective and objective sentences from movie reviews and descriptions. In addition, we automatically compiled a second much larger dataset of 200k sentences. Both datasets are freely available for research purposes.

We describe the process of building and annotating the dataset. The dataset was annotated by two annotators with Cohen's κ inter-annotator agreement equal to 0.83. In the paper, we provide a summary of the annotation guidelines used by the annotators.

We perform a series of monolingual experiments with five pre-trained BERT-like models to obtain the baseline results for the newly created Czech dataset and we are able to achieve 93.5%6 of accuracy with the XLM-R-Large model. We also perform monolingual experiments for the existing English subjectivity dataset with three models obtaining 97.28% of accuracy, which is on par with the current state-of-the-art results for this dataset. Finally, we conduct zero-shot cross-lingual subjectivity classification to verify the usability of our dataset as the cross-lingual benchmark for pre-trained multilingual models that allow transfer learning.

Our experiments confirm that we provide a dataset of relatively high quality and it can be used as an evaluation benchmark to test the ability of pre-trained models to transfer knowledge between Czech and English.

In future work, we want to focus on using the dataset to improve sentiment analysis (polarity detection) in Czech and English. Furthermore, we would like to include sentences labeled as *unsure* in the dataset, along with a detailed error analysis of the fine-tuned models.

8. Bibliographical References

Amini, I., Karimi, S., and Shakery, A. (2019). Cross-lingual subjectivity detection for resource lean languages. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 81–90, Minneapolis, USA, June. Association for Computational Linguistics.

Amplayo, R. K., Lee, K., Yeo, J., and Hwang, S.-W. (2018). Translations as additional contexts for sentence classification. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 3955–3961. AAAI Press.

Bragg, J., Cohan, A., Lo, K., and Beltagy, I. (2021). Flex: Unifying evaluation for few-shot nlp. *Advances in Neural Information Processing Systems*, 34.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (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, July. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

Esuli, A. and Sebastiani, F. (2006). Determining term subjectivity and term orientation for opinion mining. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 193–200, Trento, Italy, April. Association for Computational Linguistics.

Habernal, I., Ptáček, T., and Steinberger, J. (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, June. Association for Computational Linguistics.

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

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Hal Daumé III et al., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR, 13–18 Jul.

Khodak, M., Saunshi, N., Liang, Y., Ma, T., Stewart, B., and Arora, S. (2018). A la carte embedding: Cheap but effective induction of semantic feature vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Melbourne, Australia, July. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kocián, M., Náplava, J., Štancl, D., and Kadlec, V. (2021). Siamese bert-based model for web search relevance ranking evaluated on a new czech dataset. *arXiv preprint arXiv:2112.01810*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

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

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June. Association for Computational Linguistics.

Mihalcea, R., Banea, C., and Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983, Prague, Czech Republic, June. Association for Computational Linguistics.

Nandi, R., Maiya, G., Kamath, P., and Shekhar, S. (2021). An empirical evaluation of word embedding models for subjectivity analysis tasks. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–5. IEEE.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain, July.

Piskorski, J., Laskova, L., Marcińczuk, M., Pivovarova, L., Přibáň, P., Steinberger, J., and Yangarber, R. (2019). The second cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 63–74, Florence, Italy, August. Association for Computational Linguistics.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (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, August. Association for Computational Linguistics.

Přibáň, P. and Steinberger, J. (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, September. INCOMA Ltd.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Riloff, E., Wiebe, J., and Phillips, W. (2005). Ex-

ploiting subjectivity classification to improve information extraction. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3*, AAAI'05, page 1106–1111. AAAI Press.

Saralegi, X., San Vicente, I., and Ugarteburu, I. (2013). Cross-lingual projections vs. corpora extracted subjectivity lexicons for less-resourced languages. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 96–108. Springer.

Sido, J., Pražák, O., Přibáň, P., Pašek, J., Seják, M., and Konopík, M. (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, September. INCOMA Ltd.

Straka, M., Náplava, J., Straková, J., and Samuel, D. (2021). Robeczech: Czech roberta, a monolingual contextualized language representation model. *arXiv preprint arXiv:2105.11314*.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium, October. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Veselovská, K. and Bojar, O. (2013). Czech SubLex 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November. Association for Computational Linguistics.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, et al., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Wang, S., Fang, H., Khabsa, M., Mao, H., and Ma, H. (2021). Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*.

Wiebe, J. and Mihalcea, R. (2006). Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1065–1072, Sydney, Australia, July. Association for Computational Linguistics.

Wiebe, J. and Wilson, T. (2002). Learning to disambiguate potentially subjective expressions. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Wiebe, J. M., Bruce, R. F., and O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 246–253, College Park, Maryland, USA, June. Association for Computational Linguistics.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational linguistics*, 30(3):277–308.

Zhao, H., Lu, Z., and Poupart, P. (2015). Self-adaptive hierarchical sentence model. In *Twenty-fourth international joint conference on artificial intelligence*.

A. Appendix

A.1. Hyper-parameters

During fine-tuning, we tried a variety of hyper-parameters, we use the Adam (Kingma and Ba, 2014) optimizer with default parameters ($\beta_1 = 0.9, \beta_2 = 0.999$) and the cross-entropy loss function. We randomly shuffle training data before each epoch. We run the experiments for at most ten epochs with the linear learning rate decay (without learning rate warm-up) with the initial learning rates ranging from 2e-7 to 2e-4. The 2e-4 learning rate was used only for the Czech Electra model, when used with other models, the models started to diverge. The batch size is always set to 32 and the max length of the input sequence is 200. In Tables 9, 12, 13, 10 and 11 we report results with the used initial learning rate and a number of epochs in parentheses. The first number in brackets is the initial learning rate and the second is the number of epochs for fine-tuning.

Model	Subj-CS (cs-train) cs-test	Subj-CS-L (cs-L-train) cs-test
Czech Electra	91.85 ± 0.27 (2e-4 / 4)	91.21 ± 0.08 (2e-5 / 7)
Czert-B	92.85 ± 0.20 (2e-5 / 3)	91.79 ± 0.07* (2e-6 / 2)
RobeCzech	93.29 ± 0.19* (2e-5 / 7)	91.63 ± 0.08 (2e-6 / 2)
mBERT	91.23 ± 0.21 (2e-5 / 3)	91.14 ± 0.11 (2e-6 / 5)
XLM-R-Large	**93.56 ± 0.13** (2e-5 / 4)	**91.96 ± 0.10** (2e-6 / 9)

Table 9: Results with model hyper-parameters for Czech monolingual experiments reported as average accuracy for the testing `cs-test` data part. The * symbol denotes results containing intersection in confidence interval with the best model.

Model	CS → EN (cs-train)		CS → EN (cs-L-train)		Monolingual (en-train)
	cs-dev	en-test	en-dev	en-test	en-test
mBERT	92.11 ± 0.38	88.99 ± 0.94 (2e-5 / 3)	85.80 ± 0.89	85.53 ± 0.98 (2e-6 / 1)	95.87 ± 0.13 (2e-5 / 10)
XLM-R-Large	94.40 ± 0.36	**92.86 ± 0.44** (2e-5 / 4)	93.35 ± 0.22	**90.98 ± 0.26** (2e-7 / 1)	97.28 ± 0.07 (2e-6 / 10)

Table 10: Accuracy results with model hyper-parameters for cross-lingual experiments from Czech to English along with the results for models trained on monolingual data.

Model	Joint (cs-train + en-train)		Monolingual (cs-train)	Monolingual (en-train)
	cs-test	en-test	cs-test	en-test
mBERT	91.12 ± 0.24	95.69 ± 0.22 (2e-5 / 3)	91.23 ± 0.21 (2e-5 / 3)	95.87 ± 0.13 (2e-5 / 10)
XLM-R-Large	**93.85 ± 0.15**	**96.95 ± 0.12** (2e-6 / 10)	93.56 ± 0.13 (2e-5 / 4)	97.28 ± 0.07 (2e-6 / 10)

Table 11: Accuracy results with hyper-parameters for models jointly trained on English and Czech data along with the results for models trained on monolingual data.

Model	en-test	en-10-fold
BERT	96.55 ± 0.16 (2e-5 / 3)	96.87 ± 0.25 (2e-5 / 9)
mBERT	95.87 ± 0.13 (2e-5 / 10)	96.03 ± 0.24 (2e-5 / 5)
XLM-R-Large	**97.28 ± 0.07** (2e-6 / 10)	**97.34 ± 0.21** (2e-5 / 4)

Table 12: Results with model hyper-parameters for English monolingual experiments reported as average accuracy for the testing `en-test` and `en-10-fold` data parts.

Model	EN → CS		Monoling. (cs-train)
	en-dev	cs-test	cs-test
mBERT	95.38 ± 0.22	86.18 ± 0.33 (2e-5 / 10)	91.23 ± 0.21 (2e-5 / 3)
XLM-R-Large	97.60 ± 0.18	**90.75 ± 0.32** (2e-6 / 10)	93.56 ± 0.13 (2e-5 / 4)

Table 13: Accuracy results with model hyper-parameters for cross-lingual experiments from English to Czech along with the results for models trained on monolingual data.