MultiAspectEmo: Multilingual and Language-Agnostic Aspect-Based Sentiment Analysis

1st Joanna Szołomicka
Department of Artificial Intelligence
Wrocław University of Science and Technology
Wrocław, Poland
joanna.szolomicka@pwr.edu.pl

2nd Jan Kocoń
Department of Artificial Intelligence
Wrocław University of Science and Technology
Wrocław, Poland
jan.kocon@pwr.edu.pl

***Abstract*—The paper addresses the important problem of multilingual and language-agnostic approaches to the aspect-based sentiment analysis (ABSA) task, using modern approaches based on transformer models. We propose a new dataset based on automatic translation of the Polish AspectEmo dataset together with cross-lingual transfer of tags describing aspect polarity. The result is a MultiAspectEmo dataset translated into five other languages: English, Czech, Spanish, French and Dutch. In this paper, we also present the original TrAsp (Transformer-based Aspect Extraction and Classification) method, which is significantly better than methods from the literature in the ABSA task. In addition, we present multilingual and language-agnostic variants of this method, evaluated on the MultiAspectEmo and also the SemEval2016 datasets. We also test various language models for the ABSA task, including compressed models that give promising results while significantly reducing inference time and memory usage.**

***Index Terms*—aspect-based sentiment analysis, transformers, language-agnostic, multilingual**

I. Introduction

Sentiment analysis [1] is an extensively researched topic in natural language processing (NLP) in recent years. In most cases, entire text, for example opinion, is categorized into one of several predefined sentiment polarity classes. Such a solution has limits when the opinion covers many different topics to which the author has conflicting feelings. Aspect-based sentiment analysis (ABSA) is a subtask of sentiment analysis that allows one to classify individual aspects instead of the entire text [2], [3]. The task can be divided into aspect term extraction (ATE) and aspect polarity classification (APC). For example, in a sentence *The keyboard in this laptop is very quiet but the backlight does not work*, words *keyboard* and *backlight* are aspects. The APC task aims to determine the sentiment polarity the given aspect has. For example, 'keyboard' has a positive sentiment class and 'backlight' has a negative one. ABSA is used in practical applications to analyze consumer reviews on sold products to improve their quality and create an effective marketing strategy.

Most existing methods for ABSA problems are developed for the English language, a resource-rich language in contrast to Polish, Czech, etc. In this paper, we explored existing ABSA methods for six languages: Polish, English, Czech, French, Spanish and Dutch. We compared the performance of multilingual and language-agnostic transformer-based models taking into account their ability to transfer knowledge between languages. The experiments were performed on the Polish AspectEmo dataset, its machine translations into other languages, and the multilingual dataset for Task 5 in SemEval-2016. Moreover, we analyzed the performance of selected compressed models on the MultiAspectEmo dataset in Polish and English, taking into account the F1 score, inference time, and the number of model parameters. Our contribution is as follows: (1) we introduced a new MultiEmo dataset in six languages: Polish, English, Czech, French, Spanish and Dutch; (2) we reviewed different ABSA methods from literature (3) we compared XLMR and LaBse embedding models in end-to-end ABSA solution and showed that multilingual XLMR obtains better results on all languages; (4) we compared different compressed Transformer-based models and showed that monolingual, compressed models acquire high F1-score having much fewer parameters and faster inference time.

II. Related work

Aspect-based sentiment analysis in texts is studied mostly as APC problem. A separate topic is ATE, which is a sequence labeling task. This work focuses on the method that both extract aspects from the text and classify their sentiment polarity. ATE problem is very similar to named entity recognition (NER). In the NER task, named entities are located in the text and classified into pre-defined categories like organizations, locations, personal names, etc. ATE task also locates aspects and classifies them into two categories: aspect or non-aspect. In [4] the authors applied the NER method to the APC

This work was financed by (1) the National Science Centre, Poland, project no. 2019/33/B/HS2/02814 and 2021/41/B/ST6/04471; (2) the Polish Ministry of Education and Science, CLARIN-PL; (3) the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, CLARIN – Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19; (4) the statutory funds of the Department of Artificial Intelligence, Wroclaw University of Science and Technology.

2375-9259/22/$31.00 ©2022 European Union
DOI 10.1109/ICDMW58026.2022.00065

problem so that the predefined categories were composed of sentiment categories. All considered methods are based on the Transformer architecture, which has a predetermined maximal sequence length during training. The problem with longer texts is that they cannot be processed all at once, and prediction largely depends on the given context. The authors of [5] proposed contextual majority voting on predictions with varying sizes of left and right contexts. Paper [6] introduced Slavic BERT for the NER problem, a multilingual BERT fine-tuned on Slavic languages. Embedding models based on Transformer architecture can be divided into monolingual (pre-trained on one language), multilingual (pre-trained on multiple languages), or language-agnostic (pretrained on multiple languages in such a way that embeddings of the tokens that have the same meaning in different languages have very similar representations). Most of the monolingual models are trained on the English language because English data is the most accessible. Multilingual models were used in [7] where the authors focused on solving the problem of incorrect boundaries of named entities in the NER task for low-resource languages.

One of the issues reviewed in this work is the performance of compressed models. They have fewer parameters than non-compressed ones, but as a result, they have worse performance. The papers [8]–[10] describe methods for compressing multilingual models for the NER task. Authors of [11] proposed a compression method for the monolingual BERT model in the NER problem. They used additional data augmentation, which enhances results, so they are only slightly worse than for BERT. In [12] authors introduced two early-exit mechanisms for sequence labeling. The approach can save up to 66%~75% inference costs with minimal performance degradation.

The APC task aims to determine the sentiment polarity class of a given aspect in the text. Most existing works consider classification with three labels: positive, neutral, and negative. Authors of [13] use a pre-trained BERT model and fine-tune it on a smaller dataset for the APC task. In [14] pre-trained BERT was post-trained on unlabelled data from a target domain and fine-tuned for the APC task. Post-training allowed them to obtain better performance of the model. In [15] the authors proposed a sentence pair classification method with an auxiliary question obtaining state-of-the-art results. In [16] a graph convolutional network that uses the knowledge from the affective knowledge base SenticNet [17] was presented. The authors of [18] used the prompt-based method for ABSA, that predicts a masked word which is then mapped to the sentiment class label using a predefined mapping. AEN-BERT [19] is an Attentional Encoder Network that was introduced to improve recurrent neural networks with long-term patterns. The authors also proposed a label smoothing regularization which encourages the model to be less confident with uncertain labels like neutral. In [20] authors proposed a language-agnostic method based on Bidirectional Long Short-Term Memory (Bi-LSTM) network, which was evaluated separately on ATE and APC problems. The model was trained and tested on English, Spanish, French, Dutch, German, and Hindi. We compared our results with this method as a language-agnostic baseline.

TABLE I
DATASETS STATISTIC

Dataset	Domain	Language	Texts	Tokens	Annotations
AspectEmo	school	PL	493	47927	4496
	medicine	PL	385	49769	3595
	hotels	PL	495	70852	10030
	products	PL	488	66139	8251
	all	PL	1861	234687	26372
SemEval-2016	restaurants	EN	440	38454	2336
	restaurants	DU	400	32452	1575
	restaurants	FR	455	38067	2120
	restaurants	SP	895	48673	2504

TABLE II
MULTIASPECTEMO STATISTICS. THE FIRST ROW CONTAINS THE NUMBER OF ASPECTS IN THE POLISH ASPECTEMO. OTHER ROWS CONSIST OF THE DIFFERENCE BETWEEN THE NUMBER OF ANNOTATIONS IN THE TRANSLATED DATASET (MULTIASPECTEMO) AND THE ORIGINAL ASPECTEMO.

Language	SN	WN	N	WP	SP	AMB	ALL
PL	8046	933	6685	855	9109	744	26372
EN	114	15	191	12	24	-7	349
CZ	-5	3	85	-10	-16	-10	47
SP	-26	8	-22	1	-50	-5	-94
FR	236	32	304	32	155	17	776
DU	242	42	430	29	203	9	955

III. DATASET

A. AspectEmo Corpus

AspectEmo [21] is a linguistic corpus of consumer reviews manually annotated with aspect polarity. These reviews are taken from the PolEmo 2.0 corpus [22] and cover texts from 4 domains: school, medicine, hotels, and products. The AspectEmo corpus was annotated in a 2+1 scheme, i.e., two annotators annotated the same text, and a third annotator resolved inconsistencies. The dataset was annotated using six sentiment classes: strong negative (SN), weak negative (WN), strong positive (SP), weak positive (WP), neutral(N), and ambiguous (AMB). Annotation guidelines and inter-annotator agreement analysis are presented in [21]. The dataset is available in Inside–outside–beginning [23] format (IOB), as aspect extraction and classification is most often implemented as a sequence classification task, similar to that in proper name recognition [24]. Details of the number of elements in the set are given in Table I.

B. MultiAspectEmo Corpus

MultiAspectEmo is a Polish-language AspectEmo corpus machine translated into five languages: English, Czech, Spanish, French, and Dutch. We used the translator API DeepL, which allows the processing of texts containing XML markup. In the preprocessing step, the corpus in CCL format [25] was converted to text, and the aspect polarity annotations were transferred to text as XML tags. In the translation process into another language, the translator transferred the XML tags to the target language in the appropriate places. This process is not perfect, but much cheaper and faster than manual annotation.

TABLE III
MULTIASPECTEMO STATISTICS. THE FIRST ROW CONTAINS THE NUMBER OF ALL TOKENS FORMING ASPECTS IN THE POLISH ASPECTEMO. OTHER ROWS CONSIST OF THE DIFFERENCE BETWEEN THE NUMBER OF TOKENS FORMING ASPECTS IN THE TRANSLATED DATASET (MULTIASPECTEMO) AND THE ORIGINAL ASPECTEMO.

Language	SN	WN	N	WP	SP	AMB	ALL
PL	8049	934	6687	855	9117	744	26386
EN	2313	271	3055	227	2262	234	8362
CZ	1468	155	2256	121	1350	105	5455
SP	2503	268	2103	224	2431	206	7735
FR	4130	441	5178	378	3731	338	14196
DU	2562	316	3286	252	2590	234	9240

An interesting observation is that for other languages, the translation mechanism often proposes more aspect triggers than in the original language. Table III shows the difference between the number of aspects for the target language and the number of aspects in the original language (Polish), broken down by aspect category.

Table IV shows two examples of sentences containing original manual annotations in Polish and the results of machine translation with automatic transfer of tags to the target language. The first example is translated well, and the tags are also transferred well. The second example contains a translation error. The Polish word *kole* as a student jargon term meaning *kolokwium* (in English *colloquium*) is also the correct form of the word *koło*, which can be translated as *wheel*. However, both examples contain instances of the aspect expressed in Polish in the form of the default subject, so the annotation is done on the verb referring to the default subject. In English, the subject is obligatory, so the translator also moved the tag to the subject that appeared in the target language after the translation (PL ⇒ EN: *odbiegały* ⇒ *were (...) different*). Such situations cause the number of aspect determiners in the target language to change (see Table II).

C. SE-ABSA16

SE-ABSA16 is a corpus created for the SemEval-2016 workshop in Task 5 on the ABSA [26] task. It is a further version of the corpus originally presented at the SemEval-2014 workshop as SE-ABSA14 corpus [27]. The earlier corpus included reviews of laptops and restaurants, annotated with aspect polarity. Unlike the AspectEmo corpus, it also included aspect categories, such as. *food*. The current version contains a multilingual corpus (English, Arabic, Chinese, Dutch, French, Russian, Spanish, and Turkish) and multidomain (restaurants, laptops, mobile phones, digital cameras, hotels, museums, telecommunication). The largest number of submissions in the SemEval-2016 ABSA task was made for the domain *Restaurants*, mainly due to the low complexity of the annotation scheme (similar to AspectEmo), and annotations in 6 languages were given for this domain as well. Therefore, our study used this subset, limiting the languages common to MultiAspectEmo and SE-ABSA16: English, Spanish, French, and Dutch.

IV. EMBEDDING MODELS

In this work, we reviewed the method for the ABSA problem using different embedding models such as XLMR, LaBSE, and other BERT-based models.

A. XLMR and other BERT-based models

XLM-RoBERTa [3] is a multilingual adaptation of the monolingual RoBERTa model (Robustly Optimized BERT Pretraining Approach) [28]. It has the same architecture as BERT but has a different training setup and performs significantly better than BERT. The model is pre-trained to generate task general word embeddings that account for both the embedded token's right and left context. The model can be fine-tuned for specific tasks with low computational costs compared to pre-training. XLMR was trained on 100 languages obtaining similar results to state-of-the-art monolingual models. Other non-compressed BERT-based models that were used are:

- HerBERT [29] - monolingual BERT-based model trained on two large corpora of high-quality Polish texts.
- mDeBERTaV3 [30] - multilingual model trained on the same dataset as XLMR but with different objectives in generator-discriminator setup.

B. LaBSE

LaBSE (Language-agnostic BERT Sentence Embedding) [31] is a bidirectional dual-encoder based on transformer architecture with additive margin softmax (Fig. 1). The dual encoder consists of two identical pre-trained transformer encoders that share parameters. Encoders are pre-trained using masked language modeling objective (MLM) on monolingual data, and translation language modeling (TLM) on bilingual sequences [32]. In translation language modeling, the words are randomly masked in the source sequence and its translation. The goal of the model is to predict the masked token using context from both source and translation, so embeddings from both languages are closer in the vector space. The encoders encode source and target texts in parallel. The goal during training is to maximize the similarity between the source sequence and its translation (target) and minimize it between the source and other sequences. Authors applied additive margin softmax function $\mathcal{L}$ as a loss function [33] which is given by the formula in equation 1.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\frac{e^{\phi(x_i,y_i)-m}}{e^{\phi(x_i,y_i)-m}+\sum_{n=1,n\neq i}^{N} e^{\phi(x_i,y_i)}} \qquad (1)$$

where $\phi(x_i, y_i)$ is a dot product of source embedding x and translation y, N is the number of samples in a batch, y_i is a true translation of x_i. The margin m increases separability between translations and nearby non-translations.

C. Compressed models

Knowledge distillation [34] is a method that allows compression of a large model and retaining much of its performance. The technique entails training a smaller student

TABLE IV

EXAMPLES OF ORIGINAL ANNOTATIONS IN POLISH AND THE RESULTS OF MACHINE TRANSLATION WITH SIMULTANEOUS TRANSFER OF TAGS TO THE TARGET LANGUAGE. TYPES OF ASPECTS: **POSITIVE**, **NEGATIVE**, **NEUTRAL**. FOR SIMPLICITY, WE HAVE OMITTED DISTINGUISHING BETWEEN STRONGLY OR WEAKLY POSITIVE/NEGATIVE ASPECTS.

Lang.	Examples of better translations
PL	Wykład **prowadzi** zrozumiale jak jesteś na wszystkich to zaliczysz.
EN	The lecture **is** easy to understand as you are on all of them you will pass.
CZ	**Přednášky** jsou srozumitelné, pokud jste na všech, projdete.
SP	**Imparte** la conferencia de forma inteligible ya que en todas ellas aprobarás.
FR	**Il délivre** le cours de manière intelligible comme vous l’êtes sur toutes les épreuves, vous réussirez .
DU	**Hij geeft** de lezing begrijpelijk als je op allemaal bent je zult slagen.
Lang.	**Examples of inferior translations**
PL	**zadania** na 1 kole typowo skryptowe niestety na drugim 2 zupelnie **odbiegaly**
EN	**tasks** on the first wheel typically scripted unfortunately on the second wheel 2 **were** completely **different**
CZ	**úkoly** na prvním kole typicky skriptované bohužel na druhém kole 2 **byly** zcela **odlišné**
SP	**las tareas** en el primer círculo **eran** guiones típicos, lamentablemente en el segundo círculo 2 eran completamente diferentes
FR	les **affectations** sur le premier cercle **étaient des** scripts typiques, malheureusement sur le deuxième cercle 2 **étaient** complètement **différents**
DU	**opdrachten** op de eerste cirkel **waren** typische scripts, helaas op de tweede cirkel **waren** er 2 totaal **verschillend**

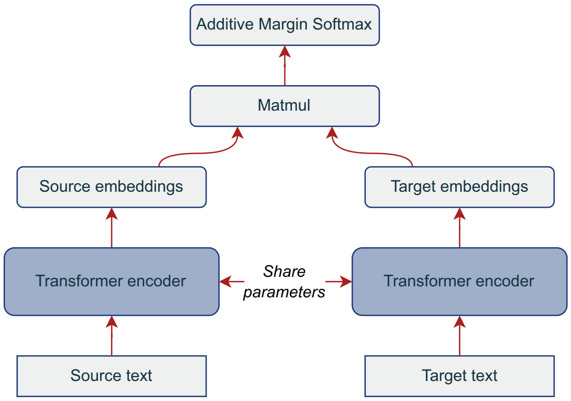

Fig. 1. LaBSE model architecture.

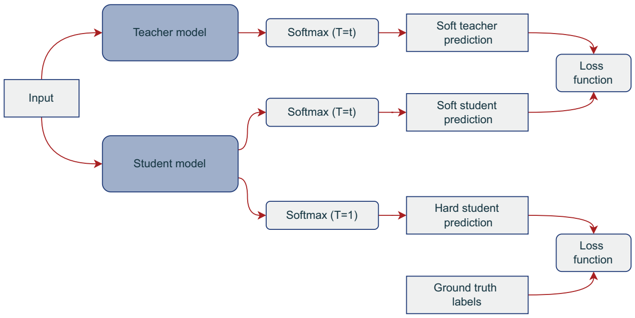

Fig. 2. Knowledge transfer from teacher to student model.

model to replicate the behavior of a larger teacher model (Fig. 2). The teacher model is trained on ground truth labels with softmax with temperature $T = t, t > 1$, generating soft predictions. The non-zero probabilities of classes other than correct represent the generalization ability of the model. The student model is trained on the teacher model’s soft predictions with softmax with $T = t$ and the ground truth labels with softmax with $T = 1$. Compressed models are especially useful for deploying on mobile devices with limited computational resources. Smaller models trained without knowledge of distillation obtain poor performance. In this work, we tested distilled models such as Polish DistilRoBERTa, miniLMv2, MobileBERT, and TinyBERT.

V. METHODS

A. Baseline

We compared our method to a language-agnostic network based on Bi-LSTM architecture [20]. The authors solve the problem of aspect-based sentiment analysis in two steps: (1) ATE and (2) APC. In the ATE solution they use dataset labeled in IOB standard [35]. To determine the polarity of a given aspect all tokens within the context window of the aspect are considered. The input to the neural network is in the form of the embeddings of each word from a single sentence and additional hand-crafted manual features.

B. TrAsp

In this work, we expanded the experiments carried out in [4], where the authors used a method for named entity recognition to solve the ABSA problem. Each token in the dataset is labeled according to the IOB standard [35]. Text embeddings from a fine-tuned pretrained transformer model are fed into three layers: linear layer, dropout, and final linear layer for classification. Figure 3 presents the architecture of the applied model (TrAsp - Transformer-based Aspect Extraction and Classification).

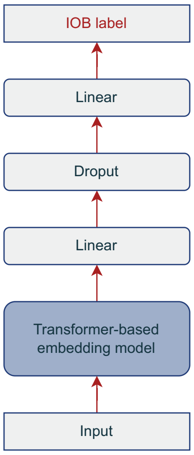

Fig. 3. TrAsp model architecture.

VI. Experiments

The experiments aimed to compare language-agnostic and multilingual models in the problem of aspect-based sentiment classification. We tested the TrAsp method with language-agnostic LaBSE [31] and multilingual XLMR-large [36] models on MultiAspectEmo and SE-ABSA16 datasets. The split for training / validation / test split had ratio 70%/15%/15%. The models were trained with maximal sequence length 256, optimizer AdamW, and cross-entropy loss function. They were trained and tested on GeForce RTX 3090 graphics card. For statistical significance analysis, all models were trained 10 times.

A. Comparison with literature

We compared the results of TrAsp with the LaBSE model and language-agnostic baseline method from literature (V-A) on the SE-ABSA16 dataset. We trained the TrAsp model separately for the ATE and APC problems. The model was trained on whole reviews instead of single sentences for a larger context. 10-fold cross-validation was performed during the evaluation of both methods. Table V shows the results of TrAsp and baseline methods for both the APC and the ATE problems. The evaluation metric for the aspect term extraction is the F1-micro score, and for the aspect polarity classification is the accuracy score. Our method obtained better results for both the ATE and APC problems. The most significant improvement can be observed for English, French, and Dutch datasets, which are two times less numerous than the Spanish. The TrAsp has a much better performance on the small datasets in comparison to the baseline method.

TABLE V
Comparison of TrAsp with LaBSE embedding model and language-agnostic baseline method on SE-ABSA16 dataset in the problem of ATE and APC.

Problem	Method	Test language			
		English	Spanish	French	Dutch
ATE (F1)	Baseline [20]	64.90	73.00	67.80	65.70
	TrAsp	78.77	74.52	79.76	72.50
APC (Acc)	Baseline [20]	83.40	87.10	74.30	81.40
	TrAsp	91.05	88.43	89.45	86.78

B. Context significance

The remaining experiments were performed using end-to-end learning where the IOB labels contained information about the aspect and the polarization class. We studied the context significance of the TrAsp model with XLMR-large embeddings trained on Polish AspectEmo in two different variations:

1) sentence: single input to the model consists of one sentence,
2) text: single input to the model is composed of a whole document.

Model trained on documents obtained better F1-micro score than on single sentences (Fig. 4). Whole text inputs provide the model access to a larger context for each predicted aspect. Models in further experiments were trained and tested on the entire document.

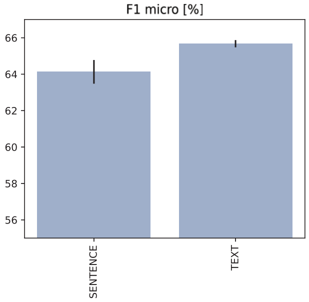

Fig. 4. Comparison of TrAsp model with XLM-RoBERTa-large embeddings trained on whole opinions and single sentences.

C. Multilingual and language-agnostic models

We compared the performance of multilingual XLMR-large and language-agnostic LaBSE using TrAsp on MultiAspectEmo and SE-ABSA16 datasets. The models were trained and evaluated using two scenarios:

- Any → Any – the model is trained and evaluated on texts in each language.
- All → Any – the model is trained on concatenated datasets in all languages and tested on texts in each language.

Table VII shows results of TrAsp with LaBSE and XLM-R on the MultiAspectEmo dataset, which is machine translated. The evaluation metrics of the models are F1-micro (exact) and F1-relaxed scores. Some aspects after translation have different or incorrect boundaries. For example, Polish aspect *słyszałem* is translated to English *I heard*, but only the word *heard* is marked as the aspect. English is not an inflectional language, and there are cases where one Polish word is translated to several English words, but the aspect is transferred to only one word. True positives are counted in the exact F1 score when the ground truth matches the prediction. In relaxed F1, true positives are an outcome when at least one word in prediction overlaps with at least one word in the ground truth. It solves the problem of incorrect boundaries in automatically translated datasets. The results show that F1-relaxed obtained better results on the translated dataset. For the model trained and tested on the original Polish AspectEmo dataset (Polish-Polish), exact and relaxed F1 scores are almost the same for both embedding models. Table VIII presents the results of the TrAsp model trained and tested on the SE-ABSA16 dataset using LaBSE and XLMR as the embedding models. The evaluation metric is the exact F1-micro score.

It can be observed that models trained on texts in language A and tested on datasets in language B obtain significantly worse results than models trained and evaluated on documents in the same language A. Models $Multi6$ and $Multi4$, which were trained in the second scenario (All → Any), achieve better results than the setup $A - B$ and $A - A$ for both embedding models on most languages. The difference between F1 scores of $Multi6$ and $Multi4$ is about 1pp. higher than in the setup $A - A$. We performed statistical tests in the following scenarios: (1) in setup $A - A$ vs $Multi - A$ both with LaBSE and XLMR, (2) setup $Multi$ with XLMR vs $Multi$ in LaBSE, (3) setup $A - A$ with XLMR vs $A - A$ in LaBSE. We checked samples normality using Shapiro-Wilk test [37]. We used student's t-test [38] with $\alpha = 0.05$. Model $Multi6$ with XLMR achieves a statistically significantly higher F1 score than model $A - A$ for English, Spanish, French, and Dutch, a lower F1 score for Polish and there is no difference for Czech. There is a statistically significant difference in F1 scores for model $Multi6$ with LaBSE tested on English, Czech, Spanish, French, and Dutch; the scores for $Multi6$ are higher than for models $A - A$. XLMR achives statistically significantly higher F1 scores for all languages in setups $A - A$ and $Multi$. There is no significant difference between the models trained on texts in different languages and evaluated on the same language in the setup $A - B$. There is a considerable decrease in performance on the machine-translated datasets compared to Polish AspectEmo, which is caused by errors in automatic translation (see Section III-B). The models trained on Polish and Czech obtained the lowest F1 scores on English, Spanish, French, and Dutch. Polish and Czech are the only ones that belong to the Slavic languages. Multilingual XLMR gains more than LaBSE even in the language setup $A - B$. It can be induced by token embeddings in XLMR being more precise as the model learns distinct representations for the tokens in each language. In LaBSE, the model aims for the embeddings of sequences with the same meaning in each language to be as close as possible in the vector space. As a result, the quality of the representation of a single token may be worse. Table VI shows the examples of the output of the models in $A - A$ scenario. It can be seen that the model predicted correctly all aspects in the Polish text and made the same mistake in all other langauges.

D. Compressed models

We analyzed the performance of TrAsp with different Transformer-based embedding models (Tab. IX). We compared two compressed models on Polish ApectEmo with multilingual XLMR and mDeBERTaV3 as well as monolingual HerBERT. Furthermore, we tested two compressed BERT models for the English language and compared them with XLMR, which achieved the best F1 score in all languages. The results show that monolingual models such as HerBERT and compressed Polish DistilRoBERTa acquire higher F1 scores than multilingual models having much fewer parameters. Polish DistilRoBERTa obtained an F1 score of 2.8 pp. less than non-compressed mDeBERTaV3 with more than three times fewer parameters and two times faster inference time.

On the other hand, the advantage of multilingual models is their linguistic universality. We evaluated a compressed, multilingual miniLMv2 which obtained an F1 score of almost 17 pp. less than XLMR, which is a huge drop. We reviewed two monolingual, compressed models for the English language, which are distilled using the BERT model. MobileBERT acquired an F1 score of 6.5 pp. less than XLMR with 2.5 faster inference time and 22.8 times fewer parameters. It is beneficial to use MobileBERT compared to BERT, which gains only 2.2 pp. more F1 score.

VII. Conclusions and future work

We created a multilingual machine-translated dataset for the ABSA problem which allowed us to compare the model's ability to transfer knowledge between languages. Such capacity enables one to train the model that can be used to process texts in different languages, including languages the model has not seen during training. The experiments showed that multilingual XLMR obtained better results than language-agnostic LaBSE in all language setups in a sequence labeling problem such as ABSA. Furthermore, a model trained on texts in all languages acquired a higher F1 score than a model trained on a dataset in one language, which shows that information about different languages complements each other. The models tested on the Polish language achieved better F1 scores than on other languages, which requires further investigation.

We observed that the models evaluated on machine-translated datasets have much lower scores than the original Polish dataset. In future work, we plan to manually correct aspects boundaries which were incorrectly transferred during translation.

References

[1] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.

[2] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intelligent systems*, vol. 28, no. 2, pp. 15–21, 2013.

[3] Y. Ma, H. Peng, and E. Cambria, "Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

TABLE VI
Examples of the outputs of TrAsp with XLMR in the $A - A$ setup. The **TEXT**$^{\mathrm{SC}}$ means the model predicted the correct aspect, **TEXT**$^{\mathrm{SC}}$ signifies the model predicted the aspect that is not in the ground truth and **TEXT** indicates the model did not predict the aspect that is the ground truth. $\mathrm{SC} \in \{\mathrm{SN}, \mathrm{WN}, \mathrm{SP}, \mathrm{WP}, \mathrm{N}, \mathrm{AMB}\}$ indicates the sentiment class of the predicted aspect.

Lang.	Examples
PL	Bardzo **sympatyczna**$^{\mathrm{SP}}$ chętnie **pomaga**$^{\mathrm{SP}}$ i **wie**$^{\mathrm{SP}}$ o co biega w tej chemii. **Zaliczenie**$^{\mathrm{SP}}$ bezproblemowe.
EN	Very **nice**$^{\mathrm{SP}}$, **willing**$^{\mathrm{SP}}$ to **help** and **knows**$^{\mathrm{SP}}$ what's going on in this chemistry. **Passing**$^{\mathrm{SP}}$ without any problems.
CZ	Velmi **přátelský**$^{\mathrm{SP}}$, **ochotný**$^{\mathrm{SP}}$ **pomoci** a **ví**$^{\mathrm{SP}}$, co se v této chemii děje. **Prochází**$^{\mathrm{SP}}$ bez problémů.
SP	Muy **amable**$^{\mathrm{SP}}$ dispuesto a **ayudar**$^{\mathrm{SP}}$ y **sabe**$^{\mathrm{SP}}$ de qué va la química. **Pasando**$^{\mathrm{SP}}$ sin problemas.
FR	Très **sympathique**$^{\mathrm{SP}}$, prêt **à**$^{\mathrm{SP}}$ **aider** et qui **sait**$^{\mathrm{SP}}$ ce qu'est la chimie. **Passage**$^{\mathrm{SP}}$ sans problème.
DU	Zeer **vriendelijk**$^{\mathrm{SP}}$ **bereid**$^{\mathrm{SP}}$ om **te helpen** en **weet**$^{\mathrm{SP}}$ waar de chemie over gaat. **Ik**$^{\mathrm{SP}}$ slaag zonder problemen.

TABLE VII
Comparison of language-agnostic LaBSE model and multilingual XLMR-large using TrAsp method on MultiAspectEmo dataset. Each model was trained and tested on texts in each language. The evaluation metric is F1-micro/F1-relaxed scores.

Language model	Train language	Test language					
		Polish	English	Czech	Spanish	French	Dutch
LaBSE	Polish	61.90/61.91	35.70/49.78	44.76/52.35	39.24/52.67	33.88/49.41	35.52/49.76
	English	55.27/56.46	45.15/55.44	44.30/51.27	40.31/52.88	36.28/52.46	39.57/51.66
	Czech	50.16/55.37	36.61/49.31	47.30/53.31	35.88/48.94	34.46/47.75	35.08/47.97
	Spanish	53.47/56.14	38.76/51.58	43.43/50.13	43.24/54.25	35.67/51.14	37.89/50.39
	French	52.65/55.88	41.80/53.12	44.16/50.82	39.14/52.55	42.60/54.08	38.87/51.34
	Dutch	53.69/56.34	41.79/52.93	43.42/50.65	39.05/52.01	35.28/51.88	43.27/54.20
	Multi6	61.53/61.63	46.07/56.44	48.43/54.60	44.75/55.60	43.75/55.33	44.53/55.49
XLM-R	Polish	65.21/65.23	38.65/54.33	47.52/55.76	41.93/56.04	35.92/52.89	38.15/53.10
	English	60.09/62.52	48.22/58.30	47.54/55.32	43.52/57.16	37.83/55.99	43.00/55.75
	Czech	54.03/61.20	41.04/55.41	50.47/56.78	40.82/55.18	38.10/53.07	39.65/54.09
	Spanish	58.19/61.58	42.48/55.40	46.89/54.68	46.66/57.35	37.51/54.17	41.26/54.43
	French	55.64/61.03	43.89/56.25	47.55/54.78	42.41/56.62	45.04/56.26	42.61/55.81
	Dutch	57.83/61.85	45.46/57.16	47.53/55.00	43.43/56.80	38.25/55.58	46.22/57.20
	Multi6	64.07/64.12	48.88/58.73	50.67/56.79	47.43/58.01	46.36/57.46	47.33/58.01

TABLE VIII
Comparison of language-agnostic LaBSE model and multilingual XLMR-large using TrAsp method on SE-ABSA16 dataset. Each model was trained and tested on texts in each language. The evaluation metric is F1-micro scores.

Language model	Train language	Test language			
		English	Spanish	French	Dutch
LaBSE	English	70.30	64.36	56.15	60.65
	Spanish	50.20	69.19	56.15	61.07
	French	52.51	67.19	66.72	58.46
	Dutch	54.09	62.99	56.8	62.98
	Multi4	74.61	72.59	70.70	68.00
XLM-R	English	74.93	71.13	62.91	66.31
	Spanish	58.86	72.97	61.26	65.13
	French	60.56	71.88	71.22	65.05
	Dutch	65.33	70.73	61.27	68.56
	Multi4	76.27	75.92	72.49	70.85

[4] J. Kocoń, J. Radom, E. Kaczmarz-Wawryk, K. Wabnic, A. Zajączkowska, and M. Zaśko-Zielińska, "AspectEmo 1.0: Multi-domain corpus of consumer reviews for aspect-based sentiment analysis," 2021, CLARIN-PL digital repository. [Online]. Available: http://hdl.handle.net/11321/849

[5] J. Luoma and S. Pyysalo, "Exploring cross-sentence contexts for named entity recognition with BERT," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 904–914. [Online]. Available: https://aclanthology.org/2020.coling-main.78

[6] M. Arkhipov, M. Trofimova, Y. Kuratov, and A. Sorokin, "Tuning multilingual transformers for language-specific named entity recognition," in *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 89–93. [Online]. Available: https://aclanthology.org/W19-3712

[7] S. Liang, L. Shou, J. Pei, M. Gong, W. Zuo, and D. Jiang, "Calibrenet: Calibration networks for multilingual sequence labeling," 2020. [Online]. Available: https://arxiv.org/abs/2011.05723

[8] H. Tsai, J. Riesa, M. Johnson, N. Arivazhagan, X. Li, and A. Archer, "Small and practical bert models for sequence labeling," 2019. [Online]. Available: https://arxiv.org/abs/1909.00100

[9] X. Wang, Y. Jiang, N. Bach, T. Wang, F. Huang, and K. Tu, "Structure-level knowledge distillation for multilingual sequence labeling," 2020. [Online]. Available: https://arxiv.org/abs/2004.03846

[10] S. Mukherjee and A. Hassan Awadallah, "XtremeDistil: Multi-stage distillation for massive multilingual models," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 2221–2234. [Online]. Available: https://aclanthology.org/2020.acl-main.202

[11] X. Zhou, X. Zhang, C. Tao, J. Chen, B. Xu, W. Wang, and J. Xiao, "Multi-grained knowledge distillation for named entity recognition," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 5704–5716. [Online]. Available: https://aclanthology.org/2021.naacl-main.454

[12] X. Li, Y. Shao, T. Sun, H. Yan, X. Qiu, and X. Huang, "Accelerating bert inference for sequence labeling via early-exit," 2021. [Online]. Available: https://arxiv.org/abs/2105.13878

TABLE IX
COMPARISON OF F1-MICRO, F1-RELAXED, INFERENCE TIME AND A NUMBER OF PARAMETERS IN TRASP MODELS WITH DIFFERENT TRANSFORMER-BASED EMBEDDING MODELS.

Language	**Language model**	**F1**	**F1-relaxed**	**Time [s]**	**Number of parameters**
Polish	HerBERT	66.46	66.46	32.65	355M
	XLMR	65.21	65.23	34.51	560M
	mDeBERTaV3	61.70	61.71	28.56	278M
	Polish DistilRoBERTa	59.89	59.89	13.47	81M
	miniLMv2	48.23	48.23	12.91	106M
English	XLMR	48.22	58.30	34.17	560M
	BERT	43.96	54.23	32.08	335M
	MobileBERT	41.74	52.16	13.74	24M
	TinyBERT	28.86	37.23	09.13	4M

[13] M. Hoang, O. A. Bihorac, and J. Rouces, "Aspect-based sentiment analysis using BERT," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Turku, Finland: Linköping University Electronic Press, Sep.–Oct. 2019, pp. 187–196. [Online]. Available: https://aclanthology.org/W19-6120

[14] H. Xu, B. Liu, L. Shu, and P. S. Yu, "Bert post-training for review reading comprehension and aspect-based sentiment analysis," 2019. [Online]. Available: https://arxiv.org/abs/1904.02232

[15] C. Sun, L. Huang, and X. Qiu, "Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence," 2019. [Online]. Available: https://arxiv.org/abs/1903.09588

[16] B. Liang, H. Su, L. Gui, E. Cambria, and R. Xu, "Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks," *Knowledge-Based Systems*, vol. 235, p. 107643, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121009059

[17] E. Cambria, Q. Liu, S. Decherchi, F. Xing, and K. Kwok, "SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 3829–3839. [Online]. Available: https://aclanthology.org/2022.lrec-1.408

[18] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," *IEEE Transactions on Affective Computing*, pp. 1–11, 2022.

[19] Y. Song, J. Wang, T. Jiang, Z. Liu, and Y. Rao, "Targeted sentiment classification with attentional encoder network," in *Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series*. Springer International Publishing, 2019, pp. 93–103.

[20] M. S. Akhtar, A. Kumar, A. Ekbal, C. Biemann, and P. Bhattacharyya, "Language-agnostic model for aspect-based sentiment analysis," in *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*. Gothenburg, Sweden: Association for Computational Linguistics, May 2019, pp. 154–164. [Online]. Available: https://aclanthology.org/W19-0413

[21] J. Kocoń, J. Radom, E. Kaczmarz-Wawryk, K. Wabnic, A. Zajączkowska, and M. Zaśko-Zielińska, "Aspectemo: multi-domain corpus of consumer reviews for aspect-based sentiment analysis," in *2021 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2021, pp. 166–173.

[22] J. Kocoń, P. Miłkowski, and M. Zaśko-Zielińska, "Multi-level sentiment analysis of polemo 2.0: Extended corpus of multi-domain consumer reviews," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 2019, pp. 980–991.

[23] H. Hamdan, P. Bellot, and F. Bechet, "Lsislif: Crf and logistic regression for opinion target extraction and sentiment polarity analysis," in *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, 2015, pp. 753–758.

[24] J. R. Curran and S. Clark, "Language independent ner using a maximum entropy tagger," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, 2003, pp. 164–167.

[25] M. Marcińczuk, J. Kocoń, and M. Janicki, "Liner2–a customizable framework for proper names recognition for polish," in *Intelligent tools for building a scientific information platform*. Springer, 2013, pp. 231–253.

[26] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. M. Jiménez-Zafra, and G. Eryiğit, "SemEval-2016 task 5: Aspect based sentiment analysis," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 19–30. [Online]. Available: https://aclanthology.org/S16-1002

[27] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 task 4: Aspect based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland: Association for Computational Linguistics, Aug. 2014, pp. 27–35. [Online]. Available: https://aclanthology.org/S14-2004

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[29] R. Mroczkowski, P. Rybak, A. Wróblewska, and I. Gawlik, "HerBERT: Efficiently pretrained transformer-based language model for Polish," in *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. Kiyv, Ukraine: Association for Computational Linguistics, Apr. 2021, pp. 1–10. [Online]. Available: https://www.aclweb.org/anthology/2021.bsnlp-1.1

[30] P. He, J. Gao, and W. Chen, "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing," 2021.

[31] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding," *CoRR*, vol. abs/2007.01852, 2020. [Online]. Available: https://arxiv.org/abs/2007.01852

[32] G. Lample and A. Conneau, "Cross-lingual language model pretraining," *CoRR*, vol. abs/1901.07291, 2019. [Online]. Available: http://arxiv.org/abs/1901.07291

[33] Y. Yang, G. H. Ábrego, S. Yuan, M. Guo, Q. Shen, D. Cer, Y. Sung, B. Strope, and R. Kurzweil, "Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax," *CoRR*, vol. abs/1902.08564, 2019. [Online]. Available: http://arxiv.org/abs/1902.08564

[34] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015. [Online]. Available: https://arxiv.org/abs/1503.02531

[35] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," 1995. [Online]. Available: https://arxiv.org/abs/cmp-lg/9505040

[36] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *CoRR*, vol. abs/1911.02116, 2019. [Online]. Available: http://arxiv.org/abs/1911.02116

[37] S. Shaphiro and M. Wilk, "An analysis of variance test for normality," *Biometrika*, vol. 52, no. 3, pp. 591–611, 1965.

[38] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.