MultiEmo: Multilingual, Multilevel, Multidomain Sentiment Analysis Corpus of Consumer Reviews

Jan Kocoń[0000−0002−7665−6896], Piotr Miłkowski[0000−0001−5201−0364], and Kamil Kanclerz[0000−0002−7375−7544]

Wrocław University of Science and Technology, Wrocław, Poland
{jan.kocon,piotr.milkowski,kamil.kanclerz}@pwr.edu.pl

Abstract. This article presents MultiEmo, a new benchmark data set for the multilingual sentiment analysis task including 11 languages. The collection contains consumer reviews from four domains: medicine, hotels, products and university. The original reviews in Polish contained 8,216 documents consisting of 57,466 sentences. The reviews were manually annotated with sentiment at the level of the whole document and at the level of a sentence (3 annotators per element). We achieved a high Positive Specific Agreement value of 0.91 for texts and 0.88 for sentences. The collection was then translated automatically into English, Chinese, Italian, Japanese, Russian, German, Spanish, French, Dutch and Portuguese. MultiEmo is publicly available under the MIT Licence. We present the results of the evaluation using the latest cross-lingual deep learning models such as XLM-RoBERTa, MultiFiT and LASER+BiLSTM. We have taken into account 3 aspects in the context of comparing the quality of the models: multilingualism, multilevel and multidomain knowledge transfer ability.

1 Introduction

Sentiment analysis has become very popular in recent years in many areas using natural language text processing. These include topics such as prediction of future events including security issues in the world [20]. There is also great interest in the analysis of consumer opinions [6], especially among product manufacturers who want to know the general reactions of customers to their products and thus improve them. Consumer reviews allow for the recognition of specific customer preferences, which facilitates good marketing decisions. With the increase in the number of reviews, especially for products sold on the global market (for which reviews are available in many languages), it is necessary to develop an effective method of multilingual analysis of the sentiment of a review, which would also be able to evaluate not only the sentiment of the entire opinion, but also its components, e.g. aspects or features of the product, whose sentiment is expressed at the level of sentences [19]. It is important that the method should also work in as many domains as possible [1].

In this work we present MultiEmo, a multilanguage benchmark corpus of consumer opinions, developed on the basis of PolEmo 2.0. The original collection was created to fill the gap in datasets annotated with sentiments for low-resource language, such as Polish. However, the results of this work show that perhaps treating Polish as a low-resource language is no longer correct (Section 7). It can certainly be said that the number of corpora annotated with sentiment for the Polish one is very small (low-resource in this domain, Section 3). Low-resource languages often provide a wealth of information related to the culture of the people who speak them. This knowledge concerns intangible cultural heritage, which allows a better understanding of the processes that have shaped a given society, its value system and traditions. These factors are important in the process of determining the sentiment of texts written by a person belonging to a particular cultural group.

MultiEmo allows building and evaluating a sentiment recognition model for both high-resource and low-resource languages at the level of the whole text, as well as single sentences, and for different domains. A high level of Positive Specific Agreement (PSA) [13] was achieved for this set, which is 0.91 for annotations at the text level and 0.81 at the sentence level. It turns out that the collection is very well suited for the evaluation of modern deep language models, especially cross-lingual ones. To the best of our knowledge, there is no other such large publicly available dataset annotated with a sentiment, allowing simultaneous evaluation of models in 3 different aspects (3M): multilingual, multilevel and multidomain.

We also present the results of classification using selected recent deep language models: XLM-RoBERTa [4], MultiFiT [8] and the proposed new combination of LASER+BiLSTM, using the Language-Agnostic SEntence Representations (LASER) [2] model to evaluate the quality of cross-lingual sentiment recognition zero-shot transfer learning task.

ID	Name	Source	Author	Subject	Size
H	hotels	tripadvisor.com	guest	hotel	3456
M	medicine	znanylekarz.pl	patient	doctor	2772
U	university	polwro.pl	student	professor	484
P	products	ceneo.pl	customer	product	504
H	hotels	ehotelarstwo.com	neutral texts		500
M	medicine	naukawpolsce.pap.pl	neutral texts		500

Table 1. The description of the review sources, with review domain, author type, subject type and domain subcorpus size (number of documents). For two domains potentially neutral texts were added as part of articles related to the domain.

2   Related Work

In recent years, the development of Transformer-based language models has led to significant improvements in cross-lingual language understanding (XLU). This

would not have been possible without an increasing number of benchmark sets, which make it possible to test the quality of new language models and compare them with existing ones. The pre-training and fine-tuning approach allows for state-of-the-art results for a large number of NLP tasks. Among the popular pre-trained models, two groups can be distinguished. The first of them are monolingual models, e.g.: BERT [7] or RoBERTa [16].The second group are multilingual models, e.g: LASER [2], XLM-RoBERTa [4], or MultiFiT [8] In this article we will focus mainly on the second group and we compare their effectiveness in aspects not only related to cross-lingual tasks, but also multidomain and multilevel. There are many benchmark data sets on which the above mentioned models are tested. In general, they can also be divided into similar groups. The following datasets can be listed in the monolingual group: GLUE [22], KLEJ [18] or CoLA [23]. In the multilingual group, the examples are: XGLUE [15] or XTREME [14].

Most of the mentioned language models support over 100 languages, e.g. LASER, mBERT, XLM, XLM-RoBERTa, fastText-RCSLS. However, there are models that are pre-trained in a much smaller number of languages, e.g. Unicoder (15 languages) or MultiFiT (7 languages). In the context of multilingual benchmark data sets, the number of supported languages is usually even smaller. The largest number of languages is XTREME (40 languages), XGLUE (19 languages) and XNLI (15 languages). However, the tasks in these datasets are mostly unique to the individual languages, i.e. they are not their translations. Additionally, there are no sets for which different levels of annotation (e.g. document level and sentence level) or other phenomena, e.g. cross-domain knowledge transfer, can be studied at the same time (i.e. on the same texts, translated into many languages). Moreover, low-resource languages are highly underrepresented in most of the sub-tasks of these benchmarks.

An important problem from the perspective of multilingual sentiment analysis is the small number of benchmark sets. None of the previously mentioned sets contain multilingual data for this task. To the best of our knowledge, there is no set for this task, which contains accurate translations of the training and test instances for many languages, additionally taking into account multidomain and multilevel aspects. We found two collections close to the one we need, but both of them did not meet our objectives. One of the existing datasets is a collection of the SemEval-2016-Task-5 [17]. One of its subtask (Out-of-domain Aspect-Based Sentiment Analysis) contains data sets for 8 languages. These are consumer reviews from different sources, but each language contains a different number of them and they are not translations of the same reviews in different languages. The next most conceptually similar set to MultiEmo is Multilanguage Tweets Corpus [9]. This collection contains 2794 tweets in Polish (1397 positive and 1397 negative), 4272 tweets in Slovenian (2312 positive and 1950 negative) and 3554 tweets in Croatian (2129 positive and 1425 negative). Then the Google Translate tool was used to translate these tweets into English. However, this data was not translated into other languages, and there were different texts within

the non-English collections. Due to a lack of data, we decided to prepare our own collection.

3   MultiEmo Sentiment Corpus

The motivation to prepare the source corpus for MultiEmo were works devoted to domain-oriented sentiment analysis, where the model is trained on annotated reviews from the source domain and tested on other [10]. A newer work on this subject describes a study on the Amazon Product Data collection [11]. However, this collection contains ratings assigned to the reviews by the authors of the texts. Additionally, annotations are assigned at the level of the entire document. The initial idea was to have a corpus of reviews that would be evaluated by the recipients, not the authors of the content. Annotations should also be assigned not only at the level of the whole document, but also at the level of individual sentences, which makes it easier to assess aspects of the opinion. The last important feature was that the collection would be multidomain in order to be able to study models in the cross-domain knowledge transfer task. Four domains presented in Table 1 were chosen to build the initial corpus. Initial set of annotation tags contained 6 different ratings: 1) Strong Positive (SP), 2) Weak Positive (WP), 3) Neutral (0), 4) Weak Negative (WN), 5) Strong Negative (SN), 6) Ambivalent (AMB). The annotators were asked not to judge the strength of sentiment when distinguishing between strong and weak categories. If the review was entirely positive or entirely negative, then it received a strong category. If the positive aspects outweighed the negative ones, then weak. If the positive and negative aspects were balanced, then the texts were marked as AMB. These rules were applied both to the entire text level and the sentence level. The final Positive Specific Agreement on a part of corpus containing 50 documents was 90% (meta) and 87% (sentence).

L	D	SN	0	AMB	SP	A
	H	91.92	99.42	78.50	91.62	89.39
	M	94.09	99.05	70.25	96.28	93.43
T	P	94.06	100.0	77.82	85.95	89.07
	S	87.50	00.00	80.78	92.52	88.32
	A	92.87	99.18	76.87	93.48	90.91
	H	93.78	88.40	65.64	93.05	89.83
	M	90.43	91.84	59.40	93.43	90.13
S	P	91.27	48.42	41.22	90.84	79.12
	S	79.21	26.56	45.48	81.39	65.68
	A	91.92	87.21	56.82	92.12	87.50

Table 2. PSA for WP/WN/AMB tags merged into one tag (AMB) at the (L)evel of (T)ext and (S)entence for the following (D)omains: (H)otels, (M)edicine, (P)roducts, (S)chool and (A)ll. Abbreviations: Strong Positive (SP), Neutral (0), Strong Negative (SN), Ambivalent (AMB).

After annotating the whole corpus, it turned out that PSA for weak categories (WP, WN, AMB) is low and does not exceed 40%. Distinguishing between the significance of positive and negative aspects was a difficult task. It was decided to merge the WP, WN and AMB categories into one AMB category. Table 2 presents the PSA value after the weak category merging procedure. After this operation, the total PSA value has increased from 83% to 91% for annotations at the text level and from 85% to 88% for annotations at the sentence level.

Type	Domain	Train	Dev	Test	SUM	Avg len
SDT	Hotels	3165	396	395	3956	773
	Medicine	2618	327	327	3272	782
	Products	387	49	48	484	756
	School	403	50	51	504	427
DOT	!Hotels	3408	427	-	3835	737
	!Medicine	3955	496	-	4451	740
	!Products	6186	774	-	6960	757
	!School	6170	772	-	6942	778
MDT	All	6573	823	820	8216	754
SDS	Hotels	19881	2485	2485	24851	92
	Medicine	18126	2265	2266	22657	111
	Products	5942	743	742	7427	98
	School	2025	253	253	2531	110
DOS	!Hotels	26093	3262	-	29355	108
	!Medicine	27848	3481	-	31329	95
	!Products	40032	5004	-	45036	101
	!School	43949	5494	-	49443	101
MDS	All	45974	5745	5747	57466	101

Table 3. The number of texts/sentences for each evaluation type in train/dev/test sets. Average length (Avg len) of line is calculated from merged set.

Table 3 shows the number of texts and sentences annotated by linguists for all evaluation types, with division into the number of elements within training, validation and test sets as well as average line length of each combined set. Finally, the corpus has been translated into 10 languages using the DeepL[1] tool: English, Chinese, Italian, Japanese, Russian, German, Spanish, French, Dutch and Portuguese. Its translations are of better quality than those generated by Microsoft Translator Hub [21]. DeepL achieves the best results when translating German texts into English or French. The semantic correctness of the translations does not guarantee the precise preservation of the sentiment associated with a given text. However, in a situation where we have limited resources and want to use information about the cultural background of authors writing in a low-resource language, machine translation is one of the best solutions. MultiEmo[2] corpus is available under the MIT Licence.

---

[1] https://www.deepl.com/

[2] https://clarin-pl.eu/dspace/handle/11321/798

4   Chosen Language Models

We have chosen XLM-RoBERTa [4] and MultiFiT [8] language models to perform analysis of sentiment recognition task and LASER [2] to test cross-lingual zero-shot transfer task capability using MultiEmo. The first model, Unsupervised Cross-lingual Representation Learning at Scale (XLM-RoBERTa), is a large multillingual language model, trained on 2.5TB of filtered CommonCrawl data, using self-supervised training techniques to achieve state-of-the-art performance in cross-lingual understanding. Unfortunately, usage of this model is a very resource-intensive process due to its complexity. The second model, Efficient Multi-lingual Language Model Fine-tuning (MultiFiT), is based on Universal Language Model Fine-Tuning (ULMFiT) [12] with number of improvements: 1) usage of SentencePiece subword tokenization instead of word-based tokenization, significantly reducing vocabulary size for morphologically rich languages, and 2) Quasi-Recurrent Neural Network (QRNN) [3] which are up to 16 times faster at train and test time comparing to long short-term memory (LSTM) neural networks due to increased parallelism. The last approach is our proposal to use LASER embeddings as an input for the neural network based on bidirectional long short-term memory (BiLSTM) architecture. During the literature review we did not find such an application directly. LASER is capable of calculating sentence embeddings for 93 languages, therefore solution prepared on one language can be used on other language without any additional training and allows performing sentiment recognition zero-shot cross-lingual transfer task. The main advantage of this multilingual approach is that a preparation of individual model for each language can be avoided. This significantly reduces the training time and memory usage. The second advantage is that it is not necessary to translate the text into each language separately. This results in a reduction of training time and the computational resources usage.

5   Multidimensional Evaluation

In order to present the multidimensional evaluation possibilities of MultiEmo, we have conducted several types of evaluation. The first three evaluation processes focused on the multilingual aspect of the sentiment corpus. The first one was to check whether models trained on LASER embeddings of texts in one language would be equally effective in sentiment analysis of texts in another language as models trained on LASER embeddings of texts in the same language as the test set. We chose 11 different languages available in MultiEmo Sentiment Corpus: Chinese, Dutch, English, French, German, Italian, Japanese, Polish, Portuguese, Russian and Spanish. The second type of evaluation aimed to check whether models trained on LASER embeddings of texts in languages other than Polish will be able to effectively analyze sentiment in texts in Polish as well as the model trained only on LASER embeddings of texts in Polish. The third evaluation focused on measuring the effectiveness of classifiers in the task of sentiment analysis in texts written in 10 different languages: Chinese, Dutch, English, French, German, Italian, Japanese, Portuguese, Russian and Spanish. We

decided to evaluate 3 different classifiers: bidirectional long short-term memory network trained on language-agnostic sentence embeddings (LASER+BiLSTM), MultiFiT and XLM-RoBERTa. The fourth evaluation focused on the multilevel aspect of the MultiEmo Sentiment Corpus. In the evaluation process we focused on checking the effectiveness of 3 classifiers (LASER+BiLSTM, MultiFiT and XLM-RoBERTa) in the sentiment recognition of single sentences. A single sentence provides far less information than a multisentence opinion. Such a small amount of information makes it difficult to correctly determine the sentiment of a review. Therefore, we decided to test the same 3 classifiers that were used in the evaluation process on text-level annotations to see if they will be equally effective in the classification of sentence-level annotations. The fifth evaluation aims to take advantage of the multidomain aspect of MultiEmo Sentiment Corpus. The sentiment of a given word often depends on the domain of the whole text. Depending on the subject of the text, the word may have positive, neutral or negative sentiment. Moreover, correct recognition of the sentiment of a text regardless of its field is an even more difficult task and requires good quality texts from many domains. During this process we evaluated 3 classifiers (LASER+BiLSTM, MultiFiT and XLM-RoBERTa) in the task of sentiment recognition in texts from a single domain. We conducted the evaluation both when the classifiers were trained on a set containing only texts from the same domain (SD) and when the training set contained texts from multiple domains (MD).

During the evaluation process we trained 30 instances of each model and then conducted evaluation on a given test set. After that we conducted statistical tests to verify the statistical significance of differences between evaluation results of each model. We decided to use independent samples t-test, as the evaluation results concerned different models. Before we conducted the test, we checked its assumptions and if any of the samples did not meet them, we used the non-parametric Mann Whitney U test. The values in bold in each table with the results of a particular evaluation, presented in Section 6 mean that a given model performed significantly better than the others. It should be mentioned that monolingual models are in fact multilingual models tuned using a single language set. In our five experiments we counted how many "cases" the model was better than the others by counting the number of occurrences of the best result in all variants in a single experiment.

Table 4 presents the average F1-score values for each of the labels as well as global F1-score, micro-AUC and macro-AUC for the MultiEmo evaluation of bidirectional long short-term memory network models trained on language-agnostic sentence embeddings. Significant differences between performance of the models trained on texts in Polish and the models trained in the same language as the test set were observed in 26 out of 70 cases (37%). The models achieved different results mainly in case of neutral and ambivalent texts, which are much more diverse than texts characterized by strong and uniform emotions, e.g. strongly positive and strongly negative.

Test lang	Train lang	SP	0	SN	AMB	F1	micro	macro
Chinese	Polish	5.62	**9.60**	2.47	**14.89**	4.17	45.05	49.23
	Chinese	**16.45**	0.72	**18.70**	0.66	**12.64**	**62.19**	52.45
Dutch	Polish	65.70	65.46	78.51	40.01	69.48	79.65	75.19
	Dutch	67.62	**73.71**	78.66	39.59	70.48	80.32	76.94
English	Polish	67.33	67.49	79.17	**42.49**	70.29	80.20	76.37
	English	69.89	71.21	77.45	35.53	70.07	80.04	76.08
French	Polish	66.02	**66.18**	78.61	**39.96**	69.54	79.74	75.45
	French	62.47	59.48	76.78	30.81	66.92	77.99	72.52
German	Polish	65.39	63.34	78.20	**38.52**	68.83	79.22	74.62
	German	**70.37**	65.07	78.76	34.81	70.43	80.29	75.48
Italian	Polish	66.18	62.80	79.08	**40.97**	69.45	79.63	75.22
	Italian	70.00	**69.77**	80.07	35.30	71.86	81.24	76.73
Japanese	Polish	**36.86**	**17.84**	**34.15**	**11.00**	**19.33**	**75.55**	**69.29**
	Japanese	3.05	0.75	21.35	0.00	12.10	60.99	50.57
Portuguese	Polish	66.59	65.13	79.53	**41.17**	70.06	80.04	75.76
	Portuguese	67.42	66.57	77.29	32.61	69.00	79.33	74.61
Russian	Polish	65.47	**64.18**	79.02	**39.22**	69.38	79.59	**75.01**
	Russian	65.46	43.54	75.43	31.19	65.43	76.95	70.56
Spanish	Polish	66.67	**65.50**	79.44	40.91	70.07	80.05	75.72
	Spanish	65.02	56.33	75.41	38.23	66.68	77.79	73.77

Table 4. Average F1-scores for the MultiEmo evaluation of LASER+BiLSTM models trained on texts in Polish and the ones trained on texts in the same language as the test set. The values in bold refer to model that achieved significantly better results than the other one. Abbreviations: Strong Positive (SP), Neutral (0), Strong Negative (SN), Ambivalent (AMB).

Table 5 shows average F1-scores for the MultiEmo evaluation of long short-term memory neural network models trained on language-agnostic sentence embeddings on the test set containing only texts in Polish. The results of models trained on texts in languages different than Polish were compared with the results of the model trained only on texts in Polish. On the basis of statistical tests described in Section 5, significant differences in model results were observed in 3 out of 70 cases (4.3%). The worst results were observed for models trained on Chinese and Japanese texts.

The MultiEmo multilingual evaluation results of different classifiers are presented in Table 6. We decided to choose three classifiers: LASER+BiLSTM, MultiFiT and XLM-RoBERTa. MultiFiT achieved the best results in 32 out of 49 cases (65%). XLM-RoBERTa outperformed other models in 38 out of 70 cases (54%). Both MultiFiT and XLM-RoBERTa obtained better results than LASER+BiLSTM in every case. MultiFiT performed better than XLM-RoBERTa in 4 out of 7 languages (57%).

6 Results

The results of the evaluation on the MultiEmo sentence-based multidomain dataset are described in Table 7. MultiFiT outperformed other models in 28 out of 28 cases (100%). XLM-RoBERTa achieved the best results in 13 out of 42 cases (31%).

Train lang.	SP	0	SN	AMB	F1	micro	macro
Polish	66.49	63.20	78.44	36.74	69.04	79.36	74.49
Chinese	27.20	1.70	21.48	1.32	30.22	53.48	47.84
Dutch	66.36	**70.85**	77.24	37.30	68.97	79.31	75.67
English	**71.44**	67.54	77.52	36.18	70.50	80.33	75.99
French	62.46	53.98	75.89	25.68	65.79	77.20	71.12
German	**71.17**	64.40	78.25	31.21	70.09	80.06	74.74
Italian	70.46	70.33	79.64	33.92	71.50	81.00	76.60
Japanese	2.83	2.46	58.83	0.00	42.06	61.37	50.61
Portuguese	63.40	57.57	74.29	29.08	65.59	77.06	71.94
Russian	63.29	37.47	74.60	26.37	63.85	75.90	68.76
Spanish	63.84	50.81	73.57	35.38	64.80	76.54	72.04

Table 5. Average F1-scores for the MultiEmo evaluation of LASER+BiLSTM models on the test set containing only texts in Polish. The values in bold refer to models that achieved significantly better results than the model trained on texts in Polish. Abbreviations: Strong Positive (SP), Neutral (0), Strong Negative (SN), Ambivalent (AMB).

Table 8 shows the evaluation results on MultiEmo single-domain and multidomain datasets. We decided to evaluate three classifiers: LASER+BiLSTM, MultiFiT and XLM-RoBERTa. In case of single domain datasets MultiFiT obtained the best results in 8 out of 16 cases (50%). XLM-RoBERTa outperformed other models in 10 out of 24 cases (42%). LASER+BiLSTM turned out to be the best in 6 out of 24 cases (25%). It outperformed other models in the review domain, achieving the best results in 5 out of 6 cases (83%). In case of multi domain evaluation XLM-RoBERTa outperformed other models in 18 out of 24 cases (75%). MultiFiT achieved the best results in 2 out of 16 cases (12.50%). The only case where LASER+BiLSTM achieved the best results were texts about products written in Japanese.

7   Conclusions and Future Work

MultiEmo service[3] with all models is available through the CLARIN-PL Language Technology Centre[4]. The source code is available on the MutliEmo GitHub page[5]. In the case of LASER+BiLSTM model evaluation few differences were found between the model trained on texts in Polish and the model trained on texts in the same language as the test set. Similarly, statistical tests showed few differences in the effectiveness of models trained on texts in different languages in the task of sentiment recognition of texts in Polish. Low values of the model average F1-scores in the case of texts in Chinese and Japanese may be related to a significantly worse quality of text translations compared to translations into languages more similar to Polish, such as English or German. On the other hand, similar values of the average F1-scores for the multilingual model in the case of

---

[3] http://ws.clarin-pl.eu/multiemo

[4] http://clarin-pl.eu/

[5] https://github.com/CLARIN-PL/multiemo

Language	Classifier	SP	0	SN	AMB	F1	micro	macro
Chinese	LASER+BiLSTM	16.45	0.72	18.70	0.66	12.64	62.19	52.45
	MultiFiT	85.81	95.02	86.78	**59.91**	83.19	88.79	87.64
	XLM-RoBERTa	**86.34**	**95.69**	**87.99**	57.13	**84.05**	**89.37**	**87.92**
Dutch	LASER+BiLSTM	67.62	73.71	78.66	39.59	70.48	80.32	76.94
	XLM-RoBERTa	**84.00**	**96.39**	**86.31**	**53.20**	**82.45**	**88.30**	**86.76**
English	LASER+BiLSTM	69.89	71.21	77.45	35.53	70.07	80.04	76.08
	XLM-RoBERTa	**85.96**	**93.76**	**88.67**	**60.47**	**84.87**	**89.91**	**88.48**
French	LASER+BiLSTM	62.47	59.48	76.78	30.81	66.92	77.99	72.52
	MultiFiT	**86.48**	**96.04**	**87.49**	**57.42**	**83.63**	**89.09**	**87.76**
	XLM-RoBERTa	83.88	95.60	86.18	51.81	81.93	87.96	86.43
German	LASER+BiLSTM	70.37	65.07	78.76	34.81	70.43	80.29	75.48
	MultiFiT	**85.85**	**96.52**	**88.21**	**60.35**	**84.22**	**89.48**	**88.28**
	XLM-RoBERTa	82.16	89.83	86.86	59.06	82.74	88.49	86.85
Italian	LASER+BiLSTM	70.00	69.77	80.07	35.30	71.86	81.24	76.73
	MultiFiT	**86.18**	**96.04**	**87.87**	57.91	83.70	89.13	87.82
	XLM-RoBERTa	85.36	93.75	87.65	**59.06**	**84.06**	**89.37**	**87.87**
Japanese	LASER+BiLSTM	3.05	0.75	21.35	0.00	12.10	60.99	50.57
	MultiFiT	83.39	**95.77**	**87.63**	58.09	82.61	88.41	87.35
	XLM-RoBERTa	**84.54**	93.60	87.41	**58.80**	**83.67**	**89.11**	**87.54**
Portuguese	LASER+BiLSTM	67.42	66.57	77.29	32.61	69.00	79.33	74.61
	XLM-RoBERTa	**85.85**	**96.87**	**86.88**	**55.69**	**83.40**	**88.93**	**87.62**
Russian	LASER+BiLSTM	65.46	43.54	75.43	31.19	65.43	76.95	70.56
	MultiFiT	**85.54**	**96.40**	86.95	**59.72**	**83.43**	**88.96**	**87.87**
	XLM-RoBERTa	82.95	90.93	**86.96**	58.94	83.22	88.81	87.10
Spanish	LASER+BiLSTM	65.02	56.33	75.41	38.23	66.68	77.79	73.77
	MultiFiT	**86.67**	95.98	**87.36**	**59.45**	**83.81**	**89.21**	**88.05**
	XLM-RoBERTa	86.28	**96.64**	87.05	56.59	83.56	89.04	87.83

Table 6. Average F1-scores for the MultiEmo evaluation of three different classifiers: LASER+BiLSTM, MultiFiT and XLM-RoBERTa. For languages not supported by MultiFiT, an evaluation was carried out for the LASER+BiLSTM and XLM-RoBERTa classifiers. The values in bold refer to model that achieved significantly better results than the other ones. Abbreviations: Strong Positive (SP), Neutral (0), Strong Negative (SN), Ambivalent (AMB).

Polish and translated texts may be related to the high similarity of the model used for machine translation and the multilingual model. The authors of DeepL do not provide information regarding this subject.

In Table 4 presenting a comparison of LASER+BiLSTM models tested on pairs of texts in Polish and the language of training data, the biggest differences are observed for the classes with the smallest number representation in the set. Analyzing F1, micro and macro results, significant differences are only for Asian languages. The results are significantly worse for these two languages than for the others. This may be due to a much smaller number of data for the LASER model for these languages, because in Table 6 the results obtained for these languages on XLM-RoBERTa and MultiFiT models are much better. Unfortunately, we do not have access to the training resources of the source models to make this clear. The results for the other languages indicate that regardless of the configuration choice, the results within a pair of two languages do not differ significantly from each other. There is a possibility that the source models (DeepL and LASER) were trained on similar data for these language pairs. On

Language	Classifier	SP	0	SN	AMB	F1	micro	macro
English	LASER+BiLSTM	**50.36**	21.97	64.33	21.65	50.95	67.30	62.17
	XLM-RoBERTa	41.34	**39.44**	**68.39**	**28.44**	**58.07**	**72.05**	**66.74**
German	LASER+BiLSTM	41.40	14.79	62.45	26.70	47.50	65.00	60.88
	MultiFiT	**76.19**	**73.17**	**77.66**	**47.25**	**72.91**	**81.94**	**79.20**
	XLM-RoBERTa	25.06	23.93	62.73	16.78	49.75	66.50	60.14
Italian	LASER+BiLSTM	44.50	27.59	64.84	24.87	49.74	66.49	62.66
	MultiFiT	**76.47**	**74.00**	**77.51**	**47.85**	**73.10**	**82.07**	**79.51**
	XLM-RoBERTa	36.20	36.26	66.76	24.32	55.71	70.47	64.87
Japanese	LASER+BiLSTM	2.46	5.30	8.82	1.56	5.14	55.24	50.93
	MultiFiT	**73.36**	**70.73**	**75.22**	**45.40**	**70.63**	**80.42**	**77.40**
	XLM-RoBERTa	45.53	44.05	69.6	30.41	59.90	73.27	68.26
Polish	LASER+BiLSTM	45.82	40.02	66.53	28.51	53.44	68.96	64.78
	XLM-RoBERTa	**52.07**	**49.81**	**73.41**	**36.59**	**64.54**	**76.36**	**71.78**
Russian	LASER+BiLSTM	46.87	4.93	61.36	18.28	46.18	64.12	59.48
	MultiFiT	**76.87**	**73.89**	**77.68**	**47.64**	**73.33**	**82.22**	**79.55**
	XLM-RoBERTa	44.82	42.67	69.74	30.07	59.89	73.26	68.17

Table 7. Average F1-scores for the evaluation on the MultiEmo sentence-based multidomain dataset. Classifiers: LASER+BiLSTM, MultiFiT, XLM-RoBERTa. For languages not supported by MultiFiT, an evaluation was carried out for the LASER + BiLSTM and XLM-RoBERTa classifiers. The values in bold refer to model that achieved significantly better results than the other ones. Abbreviations: Strong Positive (SP), Neutral (0), Strong Negative (SN), Ambivalent (AMB).

the other hand, LASER supports 93 languages and DeepL only 12. We are not able to evaluate the other languages supported by LASER, but it can be assumed that if the data representation in the source model was at a similar level as for the examined languages with a high score, we can expect equally high results for such languages. Another experiment was to compare models trained on different languages and tested only on Polish (Table 5). Aggregate results for the LASER+BiLSTM model show that models created on translations of the original set are of comparable or worse quality than the model trained on Polish. Results for some single classes turn out to be even better for models built on translations than on the model built on the original corpus. Such cases are observed for Dutch, English and German. It is possible that in the data to create source models (LASER and DeepL) for these languages there is significantly larger number of translation examples. Further work should examine the quality of the translations for individual language pairs and check the correlation between the quality of the translation and the results of models based on these translations. Table 6 shows the results of different deep multi-language models built on different MultiEmo language versions for whole texts. Similar results are available in Table 7 for models built on single sentences. The aggregate results (F1, macro, micro) show a clear superiority of XLM-RoBERTa and MultiFiT models over the zero-shot transfer learning approach. The probable cause of these differences is the use of much more texts to create DeepL, XLM-RoBERTa and MultiFiT models, compared to the LASER model. On the other hand, in the absence of a good machine translation tool, the LASER+BiLSTM model for most languages still achieves results that are at least in some business ap-

Type	Test set domain	Classifier	DE	EN	IT	JP	PL	RU
SD	Hotels	LASER+BiLSTM	75.19	74.18	77.47	7.59	68.35	61.27
		MultiFiT	**83.29**	–	**83.03**	**80.00**	–	82.03
		XLM-RoBERTa	79.49	**83.04**	73.92	77.97	**82.03**	**83.29**
	Medicine	LASER+BiLSTM	63.61	50.76	75.84	10.09	81.96	80.12
		MultiFiT	**87.77**	–	**85.32**	**85.93**	–	**87.46**
		XLM-RoBERTa	81.65	**84.71**	82.26	83.18	**82.57**	83.18
	Products	LASER+BiLSTM	58.33	39.58	**75.00**	20.83	66.67	68.75
		MultiFiT	**76.60**	–	65.96	65.96	–	72.34
		XLM-RoBERTa	72.92	**72.92**	72.92	**72.92**	**72.92**	72.92
	University	LASER+BiLSTM	**70.00**	**74.00**	**72.00**	38.00	**64.00**	**68.00**
		MultiFiT	56.00	–	66.00	52.00	–	66.00
		XLM-RoBERTa	56.00	56.00	56.00	**56.00**	56.00	56.00
MD	Hotels	LASER+BiLSTM	64.30	62.78	63.29	1.77	75.44	58.99
		MultiFiT	82.53	–	83.54	**84.56**	–	81.52
		XLM-RoBERTa	**86.84**	**85.82**	**86.08**	83.80	**87.85**	**84.81**
	Medicine	LASER+BiLSTM	75.23	78.90	75.23	2.75	83.49	51.07
		MultiFiT	87.16	–	86.07	84.10	–	85.02
		XLM-RoBERTa	**89.60**	**88.07**	**89.07**	**89.30**	**87.46**	**88.69**
	Products	LASER+BiLSTM	62.50	72.92	64.58	**95.83**	64.58	75.00
		MultiFiT	78.72	–	**87.23**	74.47	–	80.85
		XLM-RoBERTa	79.17	72.92	68.75	83.33	**77.08**	79.17
	University	LASER+BiLSTM	76.00	58.00	36.00	16.00	70.00	48.00
		MultiFiT	76.00	–	76.00	80.00	–	80.00
		XLM-RoBERTa	**78.00**	**80.00**	**92.00**	**88.00**	**80.00**	**84.00**

Table 8. Average F1-scores for the evaluation on the MultiEmo single-domain (**SD**) and multidomain (**MD**) datasets. The languages of the individual datasets: **DE** – German, **EN** – English, **IT** – Italian, **JP** – Japanese, **PL** – Polish, **RU** – Russian. Classifiers: LASER+BiLSTM, MultiFiT, XLM-RoBERTa. For languages not supported by MultiFiT, an evaluation was carried out for the LASER + BiLSTM and XLM-RoBERTa classifiers. The values in bold refer to model that achieved significantly better results than the other ones. Abbreviations: Strong Positive (SP), Neutral (0), Strong Negative (SN), Ambivalent (AMB).

plications already acceptable. The results also show that translating a text into another language using a good quality translator allows to obtain a model with results comparable to those obtained for a model built for the source language. Moreover, it has been shown that the Polish language achieves more and more satisfactory support in known SOTA tools and models, and perhaps assigning this language to the low-resource category [5] is no longer justified. Otherwise, the conclusion is that we can also get very good quality models for high-resource languages from rare resources in low-resource languages.

Table 8 shows the results of models trained on the selected domain (SD) and on all domains simultaneously (MD). The results show that in the context of domain adaptation it is not possible to clearly indicate the best model to represent a single domain (SD variants). Differences were also found in different languages within the same domain. In case one model was trained on all domains, the most domain-agnostic sentiment representation has the XLM-RoBERTa.

MultiFiT achieved the best results in the greatest number of cases. The disadvantage of this model is the small number of supported languages (only 7). XLM-RoBERTa most often achieved the second best results, except the multidomain evaluation, where it outperformed other classifiers. LASER+BiLSTM as the only zero-shot classifier obtained worse results in almost every case. In our further research, we would like to address the detailed analysis of the impact of translations on sentiment analysis. Apart from the quality of the translations as such, a relatively interesting issue seems to be a direct change in the sentiment of the text during the translation.

Acknowledgements

References

1. Al-Moslmi, T., Omar, N., Abdullah, S., Albared, M.: Approaches to cross-domain sentiment analysis: A systematic literature review. Ieee access **5**, 16173–16192 (2017)
2. Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zeroshot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics **7**, 597–610 (2019)
3. Bradbury, J., Merity, S., Xiong, C., Socher, R.: Quasi-recurrent neural networks. arXiv preprint arXiv:1611.01576 (2016)
4. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747
5. Dadas, S., Perełkiewicz, M., Poświata, R.: Evaluation of sentence representations in polish. arXiv preprint arXiv:1910.11834 (2019)
6. Day, M.Y., Lin, Y.D.: Deep learning for sentiment analysis on google play consumer review. In: 2017 IEEE international conference on information reuse and integration (IRI). pp. 382–388. IEEE (2017)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186 (2019)
8. Eisenschlos, J., Ruder, S., Czapla, P., Kadras, M., Gugger, S., Howard, J.: Multifit: Efficient multi-lingual language model fine-tuning. 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). pp. 5706–5711 (2019)
9. Galeshchuk, S., Qiu, J., Jourdan, J.: Sentiment analysis for multilingual corpora. In: Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing. pp. 120–125. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-3717

10. Glorot, X., Bordes, A., Bengio, Y.: Domain adaptation for large-scale sentiment classification: A deep learning approach. In: Proceedings of the 28th international conference on machine learning (ICML-11). pp. 513–520 (2011)

11. He, R., McAuley, J.: Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In: proceedings of the 25th international conference on world wide web. pp. 507–517. International World Wide Web Conferences Steering Committee (2016)

12. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146 (2018)

13. Hripcsak, G., Rothschild, A.S.: Technical Brief: Agreement, the F-Measure, and Reliability in Information Retrieval. JAMIA **12**(3), 296–298 (2005). https://doi.org/10.1197/jamia.M1733

14. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., Johnson, M.: Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. arXiv preprint arXiv:2003.11080 (2020)

15. Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., et al.: Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. arXiv preprint arXiv:2004.01401 (2020)

16. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

17. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S.M., Eryiğit, G.: SemEval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 19–30. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/S16-1002

18. Rybak, P., Mroczkowski, R., Tracz, J., Gawlik, I.: KLEJ: Comprehensive benchmark for polish language understanding. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1191–1201. Association for Computational Linguistics, Online (Jul 2020)

19. Shoukry, A., Rafea, A.: Sentence-level arabic sentiment analysis. In: 2012 International Conference on Collaboration Technologies and Systems (CTS). pp. 546–550. IEEE (2012)

20. Subramaniyaswamy, V., Logesh, R., Abejith, M., Umasankar, S., Umamakeswari, A.: Sentiment analysis of tweets for estimating criticality and security of events. Journal of Organizational and End User Computing (JOEUC) **29**(4), 51–71 (2017)

21. Volkart, L., Bouillon, P., Girletti, S.: Statistical vs. Neural Machine Translation: A Comparison of MTH and DeepL at Swiss Post's Language Service, pp. 145–150. Proceedings of the 40th Conference Translating and the Computer, AsLing (2018), iD: unige:111777

22. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: 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. pp. 353–355 (2018)

23. Warstadt, A., Singh, A., Bowman, S.R.: Neural network acceptability judgments. Transactions of the Association for Computational Linguistics **7**, 625–641 (2019)