applied sciences

MDPI

Article

Text-Based Emotion Recognition in English and Polish for Therapeutic Chatbot

Artur Zygadło *, **Marek Kozłowski** and **Artur Janicki**

Faculty of Electronics and Information Technology, Warsaw University of Technology, Nowowiejska 15/19, 00-665 Warsaw, Poland; Marek.Kozlowski@pw.edu.pl (M.K.); Artur.Janicki@pw.edu.pl (A.J.)
* Correspondence: zygadlo.artur@google.com

Abstract: In this article, we present the results of our experiments on sentiment and emotion recognition for English and Polish texts, aiming to work in the context of a therapeutic chatbot. We created a dedicated dataset by adding samples of neutral texts to an existing English-language emotion-labeled corpus. Next, using neural machine translation, we developed a Polish version of the English database. A bilingual, parallel corpus created in this way, named CORTEX (CORpus of Translated Emotional teXts), labeled with three sentiment polarity classes and nine emotion classes, was used for experiments on classification. We employed various classifiers: Naïve Bayes, Support Vector Machines, fastText, and BERT. The results obtained were satisfactory: we achieved the best scores for the BERT-based models, which yielded accuracy of over 90% for sentiment (3-class) classification and almost 80% for emotion (9-class) classification. We compared the results for both languages and discussed the differences. Both the accuracy and the F1-scores for Polish turned out to be slightly inferior to those for English, with the highest difference visible for BERT.

Keywords: human-machine interaction; chatbot; sentiment recognition; emotion recognition; Polish language; parallel text corpus; fastText; BERT; machine translation

check for updates

Citation: Zygadło, A.; Kozłowski, M.; Janicki, A. Text-Based Emotion Recognition in English and Polish for Therapeutic Chatbot. *Appl. Sci.* **2021**, *11*, 10146. https://doi.org/10.3390/app112110146

Academic Editor: Kyle Ke

Received: 29 September 2021
Accepted: 26 October 2021
Published: 29 October 2021

1. Introduction

Chatbots and dialogue systems are entering new areas of our lives. Quite recently, they have also been introduced to psychological and psychiatric therapies. Such a system, in order to have a therapeutic conversation with a patient, must be able to correctly recognize the patient's emotional state. In some cases, it is enough just to recognize the sentiment of the patient, i.e., to detect whether the patient's utterance has a positive or a negative emotional tinge, or carries no emotion at all. In other cases, to correctly lead the therapeutic dialogue, more detailed emotion and mood recognition must be performed.

Our long-term target is to create a therapeutic dialogue system, working in Polish, able to hold empathetic conversations with patients. Most of the existing sentiment analysis approaches for Polish are opinion-based. They work either with short texts, e.g., as in message-level Twitter sentiment polarity classification, or with longer texts, e.g., when analyzing reviews of multiple aspects of restaurants or other services or goods. However, emotion-aware dialogue agents demand a different type of dataset, one that is more empathetic and multifaceted rather than just opinion-forming.

The Polish language is under-resourced in regard to annotated empathetic texts. To fill in this gap, in our work, we made an attempt to build a Polish version of the dataset, using neural machine translation and English corpora as the source texts.

In addition, we observed that in their experiments many researchers avoid taking into account the neutral emotional state, even though neutral utterances usually prevail during conversations. Since correctly distinguishing between neutral sentiment/emotion and any other emotional state is quite important in therapy, we decided to combine emotion-labeled texts with neutral ones. Next, we ran several experiments with sentiment polarity and emotion recognition for English and Polish, comparing the results between various classifiers and both languages.

Our article is structured as follows: first, in Section 2, we briefly review the state of the art in the area of dialogue systems applied to mental health, sentiment and emotion recognition, and the existing related resources. Next, in Section 3, we describe how we created the corpora for our study. The experiments themselves are described in Section 4, followed by presentation of the results in Section 5. The article concludes with discussion of the results in Section 6 and a summary in Section 7.

2. Related Work

2.1. Conversational Agents in Mental Health

Mental disorders affect large numbers of people worldwide. To mitigate the problem of limited availability of human therapists and to develop new methods of treatment, computer-aided therapies have been designed and successfully applied in the context of mental health, including solutions based on artificial intelligence (AI) [1].

Application of AI in mental disorder therapies frequently takes the form of dialogue systems [2,3], which can be divided into two categories. The first is chatbots—virtual counselors capable of having text-based conversations with the patient, delivered, e.g., via a mobile application. Research [4–6] has shown positive impacts from chatbot-based therapy on patients with depression and anxiety. Another group of systems is the so-called embodied conversational agents (ECAs), which extend the text conversations with an animated visualization of the virtual therapist on the screen. Such systems have been applied in the treatment of depression [7] and autism spectrum disorders [8].

An important factor in human-computer interaction in therapeutic settings is the system's ability to recognize the patient's emotions and respond accordingly (*affective computing* [9]). Several affect-aware conversational systems have been developed [10,11] in the context of mental health, aiming to produce more natural and empathetic conversations; however, to the best of our knowledge, none of these were designed for the Polish language.

2.2. Text-Based Sentiment and Emotion Recognition

The long-term goal of our research is to develop a therapeutic chatbot capable of having a conversation in Polish. Such a system should not only be able to respond according to the user's intent and the topics mentioned, but its utterances should also be aligned with the user's emotional state. For this purpose, various text-based sentiment and emotion recognition methods have been developed.

The majority of historical approaches to sentiment analysis employed bag-of-words (BoW) representations and machine learning (ML) algorithms to build classifiers from textual data (e.g., utterances, opinions, reviews) with manually annotated sentiment polarity (e.g., positive, negative, neutral). Most studies focused on designing effective features to obtain better classification performance [12]. Snyder and Barzilay [13] analyzed the sentiment of multiple aspects of restaurant reviews, such as food and atmosphere. Several works have explored sentiment compositionality through careful engineering of features or polarity-shifting rules on syntactic structures [14]. Psycholinguistic features can be built using the large lexicons of word categories (LIWC) [15] that represent psycholinguistic processes (e.g., perceptual processes) and summary categories (e.g., word ratio), as well as part-of-speech categories (e.g., articles, verbs). Mohammad et al. [16] implemented diverse sentiment lexicons and a variety of handcrafted features.

Further progress toward understanding compositionality in tasks, such as sentiment detection, requires more complex datasets and more powerful ML models, such as deep-learning (DL) models. Socher et al. [17] introduced a new corpus—the Stanford Sentiment Treebank (SST), and a new approach based on DL—, the Recursive Neural Tensor Network. SST includes fine-grained sentiment labels for 215,154 phrases in the parse trees of 11,855 sentences and presents new challenges for sentiment compositionality. The proposed method, trained on the new treebank, outperformed all previous methods.

Recent works have shown that shallow neural networks can also perform well for sentiment classification. Reference [18] presents the use of fastText word embeddings [19] as representation of words to perform the task of sentiment analysis. The results showed that the proposed approach yielded better results than many classic baseline models.

Currently, BERT (Bidirectional Encoder Representations from Transformers) is reported to be the state-of-the-art language model and has achieved amazing results in many language understanding tasks [20], including sentiment recognition. In Reference [21], the authors used the pretrained BERT model and fine-tuned it for the fine-grained sentiment-classification task on the Stanford Sentiment Treebank dataset. The proposed model performed better than complicated architectures, such as paragraphVectors, or typical recursive and convolutional neural networks. However, BERT-based models also exhibit some limitations, e.g., they have large computational and memory requirements, and the black-box model characteristics make their predictions hardly interpretable.

Deep-learning approaches, such as BERT-based models, have recently achieved state-of-the-art results [22] in the *SemEval 2018* competition's task related to detecting affect in Tweets [23], with objectives ranging from emotion classification to emotion-intensity prediction. The competition dataset was annotated with 12 different emotions (incl. *no emotion*) for English, Spanish and Arabic in a multi-label manner. The most successful approaches during the competition in 2018 used combinations of sentence embeddings with features extracted from affective lexicons.

The interest in emotion analysis as part of cyclical competitions materialized one year later in *SemEval 2019*, in a task called *EmoContext* [24]. Its goal was to classify the emotion represented by a short informal dialogue utterance, also taking into account the preceding two turns of the dialogue. There were only four classes of emotion (*Happy*, *Sad*, *Angry*, and *Others*) and three different classification subtasks. Among the top systems, there were again examples of models leveraging both vector representations of sentences and emotion-related features.

Even though there are many English-language corpora of emotional texts, for example, References [13,17,25], only a few relevant resources exist for Polish. Until recently, the majority of approaches to sentiment analysis in Polish were based on lexicons, such as *plWordNet 4.0 Emo* [26,27] or the *Nencki Affective Word List (NAWL)* [28]. In recent years, several sentiment-labeled corpora have also been created. One is *PolEmo* [29], a corpus of consumer reviews from four domains: medicine, hotels, products, and schools. It contains 8216 reviews having 57,466 sentences. The corpus is labeled, both on review and sentence level, with four polarity classes: positive, neutral, negative, and ambiguous. Another example is the HateSpeech corpus [30], the current version of which contains over 2000 manually annotated posts crawled from the Polish public web. The posts contain various types and degrees of offensive language, expressed toward minorities (e.g., ethnic, racial). However, to the best of our knowledge, no emotion-labeled corpora exist for Polish. In addition, the current corpora contain no dialogue phrases. Our work aims to fill these gaps by creating a new corpus, suitable for the context of a chatbot.

2.3. Dialogue Corpora

To develop an affect-aware dialogue system, relevant conversational datasets are required. Most of the existing dialogue corpora are either domain-specific and task-oriented [31,32] or collected without full control over the content, e.g., from social media [33,34], and, therefore, are generally inappropriate in a therapeutic setting. There are, however, examples of emotionally-grounded conversational datasets for English, such as EmpatheticDialogues [35] and DailyDialog [36], which are labeled with emotions at the dialogue and utterance level, respectively.

The DailyDialog dataset [36] consists of daily conversations obtained through crawling websites for English learners. In total, it contains 13k dialogues, manually labeled with dialogue acts and emotions at the utterance level. The emotion-labeling scheme distinguishes six basic emotions: anger, disgust, fear, happiness, sadness, and surprise.

The emotion distribution in the dataset is highly imbalanced, with "happiness" being more than 10 times more frequent than the other emotions. There are also a large number of utterances (83% of the entire dataset) marked as representing no emotion.

The authors of Reference [35] propose a new benchmark for empathetic dialogue generation and EmphatheticDialogues (ED) itself—a novel dataset with about 25k personal dialogues. Each dialogue is grounded in a specific situation where the speaker was feeling a given emotion, with the listener responding actively. The resource consists of crowdsourced one-on-one conversations, and covers a large set of emotions in a balanced way. This dataset is larger and contains a more extensive set of emotions than many similar emotion-prediction datasets from other text domains. The authors' experiments show that models built on this dataset are considered more empathetic by human evaluators, compared to models merely trained on large-scale Internet-crawled opinion-oriented data.

To the best of our knowledge, no similar dialogue corpora exist for Polish.

3. Materials and Methods

We decided to create a corpus for emotion recognition from two sources: the EmpatheticDialogues and DailyDialog datasets, previously presented in Section 2.3. Within this research, following the classification experiment described in Reference [35], we frame the problem of emotion recognition as a single dialogue turn classification. Therefore, it was sufficient for creation of the corpus for classification to take just one utterance from each of the dialogues.

3.1. Extending the Corpus with Neutral Texts

The utterances in the EmpatheticDialogues corpus were collected by providing the dialogue participants with a *prompt* sentence, together with a *context* label, representing one of 32 emotional groundings in which several dialogue turns were then produced, starting from the prompt or a slightly modified version of it. In most dialogues, the emotional grounding is reflected in the first dialogue turn, but this is not always the case. Bearing that in mind, we decided to build our corpus for emotion recognition based on the prompt sentences rather than on the dialogue content itself.

Considering the planned future usage of the developed emotion detector in the context of a therapeutic chatbot, it was necessary to include neutral utterances in the classification corpus. Unfortunately, there were no examples of labeled neutral sentences in the EmpatheticDialogues dataset. Therefore, for this purpose, sentences from the DailyDialog corpus were used. For each of the experiments conducted, neutral ("no emotion") sentences were sampled from DailyDialog data, avoiding duplicated entries, equal in number to the mean count of all the other classes in the experiment. The original corpora were split into training, validation, and test subsets by their authors, and we decided to retain this division in the resulting dataset.

3.2. Creating the Polish Corpus Using Machine Translation

Since we aim to create a sentiment and emotion classifier for a chatbot working in Polish, we faced the problem of lack of relevant dialogue datasets for the Polish language. As a cheap and fast solution, we proposed to take advantage of the available resources for English and obtain the desired Polish sentences via machine translation (MT). Such an approach has already turned out to be successful, e.g., in creating a corpus for virtual assistants [37].

We tried several available MT solutions, translating from English into Polish and observed the correctness of the translations. Eventually we chose the Google Translation API (https://cloud.google.com/translate, accessed on 21 May 2021), a neural MT system, as it gave the highest level of correctness among the tested MT systems.

The whole process of creating our corpus is shown in Figure 1. Eventually we created a parallel bilingual (English and Polish) corpus of emotional texts, designed to serve experiments on sentiment and emotion recognition. We named it CORTEX, or CORpus of Translated Emotional teXts.

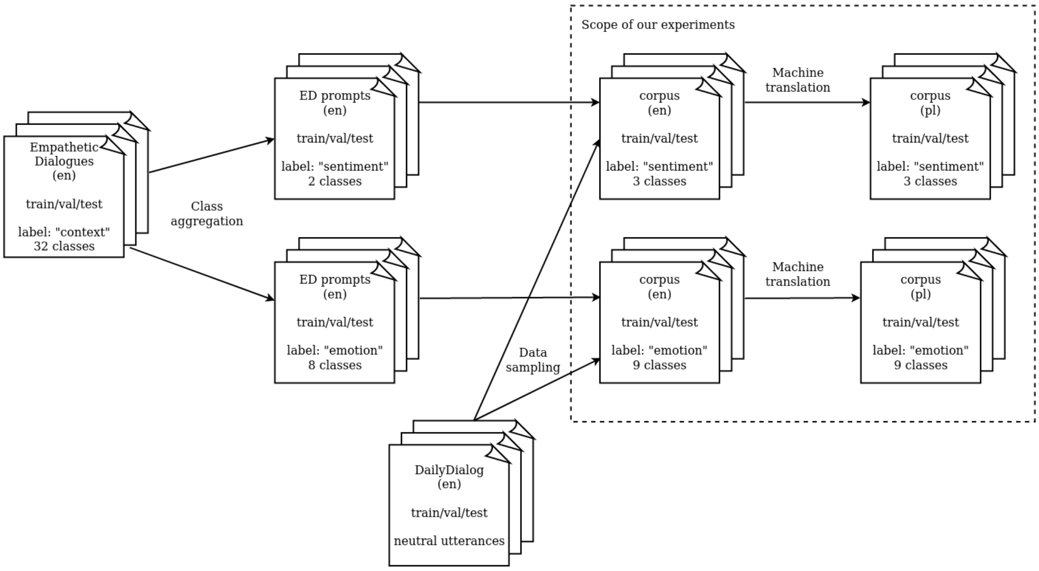
Scope of our experiments
Empathetic Dialogues (en) train/val/test label: "context" 32 classes
Class aggregation
ED prompts (en) train/val/test label: "sentiment" 2 classes
ED prompts (en) train/val/test label: "emotion" 8 classes
corpus (en) train/val/test label: "sentiment" 3 classes
corpus (en) train/val/test label: "emotion" 9 classes
Machine translation
corpus (pl) train/val/test label: "sentiment" 3 classes
Machine translation
corpus (pl) train/val/test label: "emotion" 9 classes
Data sampling
DailyDialog (en) train/val/test neutral utterances

Figure 1. Process of creating corpora for emotion recognition.

4. Experiments

4.1. Models

We evaluated several approaches to emotion recognition, both simpler and more complex ones. The baseline models were the multinomial Naïve Bayes and linear Support Vector Machine classifiers trained on top of BoW representation applied to token bigrams. Both algorithms are available within the *scikit-learn* (https://scikit-learn.org, accessed on 21 May 2021) framework. Another model was based on the fastText algorithm, obtained from pretrained word embeddings (300-dimensional variant) available in the *fasttext* library (https://fasttext.cc, accessed on 21 May 2021), both for English and Polish, and also using token bigrams. For other fastText hyperparameters, we used their default values, including training for 5 epochs and a learning rate of 0.1.

The most complex approach was to fine-tune pretrained BERT$_{\text{BASE}}$ models for English and Polish (https://huggingface.co/dkleczek/bert-base-polish-uncased-v1, accessed on 21 May 2021) (the uncased variants) to the task of sequence classification. We performed fine-tuning using the AdamW optimizer with a linear learning rate decay starting from the maximum value of learning rate of 2×10^{-5}, preceded by a warm-up for 10% of steps. We trained the models for 4 epochs with an effective batch size of 24, and we selected the best model for each experiment based on the validation metrics obtained after each training epoch. The code for training BERT was developed with the *HuggingFace Transformers* library [38].

4.2. Classes of Emotions

The corpus created as a combination of EmpatheticDialogues and DailyDialog utterances contained 32 classes of emotions plus the neutral class. This number seemed unnecessarily high considering the context of a therapeutic chatbot.

For the purpose of developing emotion classification models, we introduced new levels of class aggregation. First, we excluded some of the original emotion labels (anxious, surprised, impressed, nostalgic, sentimental, anticipating), which proved to be difficult to assign, as the utterances represented a given emotion both in positive and negative situations. Such ambiguous emotions might introduce noise to the training process. Next, we grouped similar emotion classes (see Table 1), taking into account the original papers that were the inspiration for emotion inventory used in Reference [35]. This led to two experimental setups—*sentiment* (3 classes) and *emotion* (9 classes) classification.

Table 1. Class aggregations and corpus statistics (# sentences) for *sentiment* and *emotion* setups.

Sentiment	# Sentences Train/Val/Test	Emotion	# Sentences Train/Val/Test	Classes Used in Reference [35]
positive	5985/807/826	happiness	2391/310/328	excited/joyful/grateful/content
		confidence	1705/237/222	confident/prepared/hopeful
		other positive	1889/260/276	proud/trusting/caring/faithful
negative	8314/1133/1080	sadness	2267/322/289	sad/lonely/disappointed/devastated
		anger	1804/243/226	angry/annoyed/furious
		fear	1565/212/207	afraid/terrified/apprehensive
		guilt	1576/215/199	embarrassed/ashamed/guilty
		other negative	1102/141/159	jealous/disgusted
neutral	7149/970/953	neutral	1787/242/238	no emotion
total	21,448/2910/2859	**total**	16,086/2182/2144	

5. Results

We evaluated four different models for *sentiment* and *emotion* recognition, for each of the languages, using the developed CORTEX dataset. The numbers of sentences in individual subsets (train/val/test) are displayed in Table 1. In each experiment, we measured the values of accuracy and support-weighted F1-score (see Table 2). We conducted statistical analyses using the Wilson score interval, for the confidence level set to 90%. We assessed the confidence intervals for F1-score based on the confidence intervals for precision and recall. We also generated confusion matrices, allowing for a more detailed analysis of the results, including the models' mistakes.

Table 2. Results of experiments (in percentages) on sentiment and emotion recognition for English and Polish versions of the corpus for test subset. Confidence intervals given for confidence level 90%.

Experiment	Metric	Language	Classifier			
			NB	SVM	FT	BERT
Sentiment (3-class)	Accuracy	en	85.41 ± 1.08	86.99 ± 1.03	88.07 ± 0.99	93.74 ± 0.74
		pl	83.00 ± 1.15	84.89 ± 1.10	85.59 ± 1.08	92.24 ± 0.82
	F1-score	en	85.44 ± 1.05	86.89 ± 1.02	88.00 ± 0.97	93.75 ± 0.71
		pl	83.08 ± 1.12	84.75 ± 1.08	85.43 ± 1.06	92.26 ± 0.79
Emotion (9-class)	Accuracy	en	63.29 ± 1.71	65.11 ± 1.69	68.42 ± 1.65	78.96 ± 1.44
		pl	62.27 ± 1.72	63.43 ± 1.71	65.25 ± 1.69	75.19 ± 1.53
	F1-score	en	63.06 ± 1.70	64.86 ± 1.67	68.33 ± 1.63	79.08 ± 1.40
		pl	62.06 ± 1.71	63.11 ± 1.70	65.01 ± 1.67	75.15 ± 1.49

As seen in the table above, in both experimental setups (*sentiment* and *emotion*), BERT models outperformed the simpler methods, reaching over 90% accuracy for sentiment classification and almost 80% for emotion classification. The F1-score yielded similar values. The results for the test subset were usually slightly inferior to those for the validation subset; however, the difference was not high. The results obtained for Polish were by a few relative % worse than for English, especially in the case of BERT for *emotion* recognition. Unsurprisingly, for the 3-class scenario, evaluation metrics reached much higher values than for the 9-class scenario. The baseline models achieved visibly worse results, with SVM being slightly better than Naïve Bayes. These classifiers were based on the BoW representation, and, therefore, were not able to express semantic similarity between tokens in a proper way.

We observed a slight improvement for the fastText-based classifier applied to token bigrams. In this algorithm, the embeddings were created without the context of the entire input sequence, and this might be why BERT outperformed it by a large margin.

The scores for the Polish corpora were worse than for their English counterparts. This difference was significant for the more complex models. Apart from the degraded quality of the translated sentences, this might also have been caused by the quality of the underlying pretrained embeddings. Nevertheless, as presented in Table 3, both English and Polish models learned to distinguish neutral sentences from emotional ones (the F1-score for the neutral class was around 97%). Positive and negative polarity predictions reached between 88% and 92%, with slightly higher per-class scores for negative polarity. It is worth noting that only the BERT$_{\text{BASE}}$ (uncased) variants were evaluated, while plenty of other contextual embedding models are available.

Table 3. Evaluation metrics (in percentages) for sentiment classification with BERT model for test subsets, for both languages. Confidence intervals given for confidence level 90%.

Language	Sentiment	Precision	Recall	F1-Score
English	positive	89.49 ± 0.94	92.74 ± 0.80	91.08 ± 0.87
	negative	93.54 ± 0.75	91.20 ± 0.87	92.36 ± 0.81
	neutral	97.79 ± 0.45	97.48 ± 0.48	97.64 ± 0.47
Polish	positive	87.80 ± 1.00	88.86 ± 0.96	88.33 ± 0.98
	negative	90.68 ± 0.89	91.94 ± 0.83	91.31 ± 0.86
	neutral	98.06 ± 0.42	95.49 ± 0.64	96.76 ± 0.53

6. Discussion

The envisioned study objectives have been met: we have created and tested a sentiment (3-class) and emotion (9-class) text-based classification engine for a therapeutic dialogue system, working in Polish. To achieve this, we had to create our own emotion-labeled corpus, which we generated using a neural MT system and two source English corpora. For sentiment and emotion recognition, we employed the state-of-the-art deep-learning classifier based on the BERT model, which outperformed the classic models, such as Naïve Bayes or Support Vector Machines.

We analyzed the misclassifications made by the best model (BERT) in more detail by looking at the examples related to the highest values from the confusion matrix (Table 4), especially the cases when a positive emotion label was confused with a negative emotion prediction and vice versa. The most problematic class was *other_positive*, as it was quite frequently predicted for sentences labeled with negative emotions, such as *anger*, *sadness*, and *other_negative*. The models for both languages did well in distinguishing between neutral and emotional texts; we obtained high F1-scores for the neutral class: 97.6% and 96.8% for English and Polish, respectively.

Table 4. Confusion matrix for emotion classification in Polish with BERT model for test subset. Confidence intervals given for confidence level 90%.

Emotion	Happin.	Conf.	o_pos	Anger	Fear	Sadness	Guilt	o_neg	neutral	F1-Score [%]
happiness	**247**	15	27	4	5	10	8	5	7	75.08% ± 1.53
confidence	24	**158**	15	2	7	7	3	2	4	74.70% ± 1.54
other_pos	39	16	**180**	5	6	8	10	10	2	67.04% ± 1.67
anger	3	0	7	**159**	12	13	10	18	4	70.04% ± 1.62
fear	0	4	4	8	**171**	10	7	3	0	80.28% ± 1.41
sadness	10	0	8	17	9	**215**	11	14	5	75.84% ± 1.52
guilt	0	3	7	15	2	10	**148**	13	1	73.45% ± 1.56
other_neg	3	1	11	18	3	5	7	**111**	0	66.07% ± 1.68
neutral	4	4	2	0	4	0	0	1	**223**	92.15% ± 0.95

We can explain some of the failed predictions as errors in the translation, others by prompt difficulty, e.g., the emotion was not reflected in the prompt itself (example: *I have some friends who are traveling all over Europe* taken from a dialogue labeled *jealous*) or multiple emotions were present in the utterance (example: *I recently said goodbye to a good friend for a* *while. I love her!*). Sometimes the label itself was just wrong, e.g., some of the neutral texts from DailyDialog seemed to be missing emotional labels (*I'm happy with that price*—labeled with *no emotion*—for which the model predicted *happiness*).

One of the limitations of our study is the accuracy of employed MT. Translation errors are the inevitable cost of our fast method of creating an emotion-labeled corpus for a new language. To assess the level of this inaccuracy, we manually verified a sample of our corpus. We found that about 10% of cases contained minor translation mistakes. Nevertheless, we observed that only about one-fourth of these might have an impact on the emotion category.

Considering how the dataset was obtained (machine translation from English, noisy labels), we consider the experiment results satisfactory. In the future, we plan to manually go through the developed corpus, fix the translation errors and the label mismatch where necessary, and check whether this improves the performance of our emotion-classification models.

7. Conclusions

In this article, we presented the results of our experiments on sentiment polarity and emotion recognition for English and Polish texts, aiming to work in the context of a therapeutic chatbot. We extended the existing language resources by adding samples of neutral texts to an existing English corpus. Next, we created a Polish version of the English database using neural machine translation. We used the corpus created in this way, which we named CORTEX, for experiments on sentiment and emotion classification. To show statistical significance, we calculated the Wilson score interval for each evaluation metric.

The results obtained were satisfactory: the best scores were achieved for the BERT-based classifiers, where accuracy of over 90% was achieved for sentiment (3-class) classification and almost 80% for emotion (9-class) classification. The results for Polish always turned out inferior to those for English, which might be caused either by imperfections in the MT process, or by the nature of the Polish language itself, as it is characterized by a more complex grammar and morphology. Exact research on this topic will be the subject of future work.

Our novel contributions presented in this article are as follows:

- From existing resources, we created a new dataset containing empathetic utterances in English, which were annotated with nine emotion classes, including neutral texts.
- Using neural machine translation, we created a Polish version of the above database, thus filling a gap in text resources for the Polish language. The two language versions of the database formed a new parallel corpus, named CORTEX.
- We ran a series of experiments with sentiment polarity and emotion classification, establishing that the BERT-based classifier is currently the best method. Thus, we set a baseline for potential future researchers.
- We showed the difference in classification efficacy for English and Polish and discussed possible explanations for this.

We made CORTEX, the developed dataset, available to the research community at https://github.com/azygadlo/CORTEX, accessed on 21 May 2021 and encourage researchers to use it for future experiments. We believe that this will help in designing better, more empathetic chatbots and dialogue systems, both for English and Polish. We also strongly encourage the creation of new versions of our database by extending it with next language versions.

Author Contributions: Conceptualization, A.Z., M.K. and A.J.; methodology, A.Z., M.K. and A.J.; software, A.Z.; validation, A.Z.; investigation, A.Z., M.K. and A.J.; data curation, A.Z.; writing—original draft preparation, A.Z., M.K. and A.J.; writing—review and editing, A.Z., M.K. and A.J.; visualization, A.Z.; supervision, A.J.; project administration, A.J.; funding acquisition, A.Z., M.K. and A.J. All authors have read and agreed to the published version of the manuscript.

Funding: Research was funded by the Center for Priority Research Area Artificial Intelligence and Robotics of the Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) program.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: CORTEX is freely available at https://github.com/azygadlo/CORTEX, accessed on 21 May 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Luxton, D. *Artificial Intelligence in Behavioral and Mental Health Care*; Academic Press: Cambridge, MA, USA, 2015; pp. 1–293. [CrossRef]
2. Abd-alrazaq, A.A.; Alajlani, M.; Alalwan, A.A.; Bewick, B.M.; Gardner, P.; Househ, M. An overview of the features of chatbots in mental health: A scoping review. *Int. J. Med. Inform.* **2019**, *132*, 103978. [CrossRef] [PubMed]
3. Laranjo, L.; Dunn, A.; Tong, H.L.; Kocaballi, A.; Chen, J.A.; Bashir, R.; Surian, D.; Gallego, B.; Magrabi, F.; Lau, A.; et al. Conversational agents in healthcare: A systematic review. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 1248–1258. [CrossRef] [PubMed]
4. Fitzpatrick, K.K.; Darcy, A.; Vierhile, M. Delivering Cognitive Behavior Therapy to Young Adults with Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Ment. Health* **2017**, *4*, e19. [CrossRef] [PubMed]
5. Fulmer, R.; Joerin, A.; Gentile, B.; Lakerink, L.; Rauws, M. Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. *JMIR Ment. Health* **2018**, *5*, e64. [CrossRef] [PubMed]
6. Inkster, B.; Sarda, S.; Subramanian, V. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR Mhealth Uhealth* **2018**, *6*, e12106. [CrossRef] [PubMed]
7. Ring, L.; Bickmore, T.; Pedrelli, P. An Affectively Aware Virtual Therapist for Depression Counseling. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2016) Workshop on Computing and Mental Health, San Jose, CA, USA, 7–12 May 2016; p. 01951-12.
8. Tanaka, H.; Negoro, H.; Iwasaka, H.; Nakamura, S. Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PLoS ONE* **2017**, *12*, e0182151. [CrossRef] [PubMed]
9. Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 1997.
10. Ghandeharioun, A.; McDuff, D.; Czerwinski, M.; Rowan, K. EMMA: An Emotion-Aware Wellbeing Chatbot. In Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), Cambridge, UK, 3–6 September 2019; pp. 1–7. [CrossRef]
11. Miner, A.; Chow, A.; Adler, S.; Zaitsev, I.; Tero, P.; Darcy, A.; Paepcke, A. Conversational Agents and Mental Health: Theory-Informed Assessment of Language and Affect. In Proceedings of the 4th International Conference on Human Agent Interaction (HAI 2016), Singapore, 4–7 October 2016; pp. 123–130. [CrossRef]
12. Pang, B.; Lee, L. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* **2008**, *2*, 1–135. [CrossRef]
13. Snyder, B.; Barzilay, R. Multiple aspect ranking using the good grief algorithm. In Proceedings of the Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference, Rochester, NY, USA, 22–27 April 2007; pp. 300–307.
14. Nakagawa, T.; Inui, K.; Kurohashi, S. Dependency tree-based sentiment classification using CRFs with hidden variables. In Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, CA, USA, 2–4 June 2010; pp. 786–794.
15. Pennebaker, J.W.; Francis, M.E.; Booth, R.J. Linguistic inquiry and word count: LIWC 2001. *Mahway Lawrence Erlbaum Assoc.* **2001**, *71*, 2001.
16. Mohammad, S.M.; Turney, P.D. Crowdsourcing a word–emotion association lexicon. *Comput. Intell.* **2013**, *29*, 436–465. [CrossRef]
17. Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C.D.; Ng, A.Y.; Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1631–1642.
18. Santos, I.; Nedjah, N.; de Macedo Mourelle, L. Sentiment analysis using convolutional neural network with fastText embeddings. In Proceedings of the 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI), Arequipa, Peru, 8–10 November 2017; pp. 1–5.
19. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
20. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
21. Munikar, M.; Shakya, S.; Shrestha, A. Fine-grained sentiment classification using BERT. In Proceedings of the 2019 Artificial Intelligence for Transforming Business and Society (AITB), Kathmandu, Nepal, 5 November 2019; Volume 1, pp. 1–5.

22. Alhuzali, H.; Ananiadou, S. SpanEmo: Casting Multi-label Emotion Classification as Span-prediction. *arXiv* **2021**, arXiv:2101.10038.
23. Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; Kiritchenko, S. SemEval-2018 Task 1: Affect in Tweets. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018; Association for Computational Linguistics: New Orleans, LA, USA, 2018; pp. 1–17. [CrossRef]
24. Chatterjee, A.; Narahari, K.; Joshi, M.; Agrawal, P. SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text. In Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019), Minneapolis, MN, USA, 6–7 June 2019; ACL: Minneapolis, MN, USA, 2019; pp. 39–48. [CrossRef]
25. Pang, B.; Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv* **2004**, arXiv:0409058.
26. Zaśko-Zielińska, M.; Piasecki, M.; Szpakowicz, S. A Large Wordnet-based Sentiment Lexicon for Polish. In Proceedings of the International Conference Recent Advances in Natural Language Processing, Hissar, Bulgaria, 7–9 September 2015; Incoma Ltd.: Shoumen, Bulgaria; Hissar, Bulgaria, 2015; pp. 721–730.
27. Kocoń, J.; Janz, A.; Piasecki, M. Classifier-based Polarity Propagation in a WordNet. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; European Language Resources Association (ELRA): Miyazaki, Japan, 2018.
28. Riegel, M.; Wierzba, M.; Wypych, M.; Żurawski, L.; Jednoróg, K.; Grabowska, A.; Marchewka, A. Nencki Affective Word List (NAWL): The cultural adaptation of the Berlin Affective Word List–Reloaded (BAWL-R) for Polish. *Behav. Res. Methods* **2015**, *47*, 1222–1236. [CrossRef] [PubMed]
29. Kocoń, J.; Miłkowski, P.; Zaśko-Zielińska, M. 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), Hong Kong, China, 3–4 November 2019; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 980–991. [CrossRef]
30. Troszyński, M.; Wawer, A. Czy komputer rozpozna hejtera? Wykorzystanie uczenia maszynowego (ML) w jakościowej analizie danych. [Can a Computer Recognize Hate Speech? Machine Learning (ML) in Qualitative Data Analysis]. *PrzegląD Socjol. Jakościowej* **2017**, *XIII*, 62–80.
31. Budzianowski, P.; Wen, T.H.; Tseng, B.H.; Casanueva, I.; Ultes, S.; Ramadan, O.; Gašić, M. MultiWOZ—A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2–4 November 2018; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 5016–5026. [CrossRef]
32. Hemphill, C.T.; Godfrey, J.J.; Doddington, G.R. The ATIS Spoken Language Systems Pilot Corpus. In Proceedings of the Speech and Natural Language: Proceedings of a Workshop, Hidden Valley, PA, USA, 24–27 June 1990.
33. Henderson, M.; Budzianowski, P.; Casanueva, I.; Coope, S.; Gerz, D.; Kumar, G.; Mrkšić, N.; Spithourakis, G.; Su, P.H.; Vulić, I.; et al. A Repository of Conversational Datasets. *arXiv* **2019**, arXiv:1904.06472.
34. Ritter, A.; Cherry, C.; Dolan, W.B. Data-Driven Response Generation in Social Media. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11), Edinburgh, UK, 27–31 July 2011; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 583–593.
35. Rashkin, H.; Smith, E.M.; Li, M.; Boureau, Y.L. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv* **2018**, arXiv:1811.00207.
36. Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; Niu, S. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Taipei, Taiwan, 27 November–1 December 2017; Asian Federation of Natural Language Processing: Taipei, Taiwan, 2017; pp. 986–995.
37. Sowański, M.; Janicki, A. Leyzer: A Dataset for Multilingual Virtual Assistants. In *Lecture Notes in Computer Science, Proceedings of the Conference on Text, Speech, and Dialogue (TSD2020), Brno, Czech Republic, 8–11 September 2020*; Sojka, P., Kopeček, I., Pala, K., Horák, A., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 477–486.
38. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.