Enhanced Emotion and Sentiment Recognition for Empathetic Dialogue System Using Big Data and Deep Learning Methods*

Marek Kozłowski[1][0000−0002−6313−8387], Karolina Gabor-Siatkowska[1][0000−0001−7458−5003], Izabela Stefaniak[2][0000−0001−7332−6157], Marcin Sowański[1][0000−0002−9360−1395], and Artur Janicki[1][0000−0002−9937−4402]

[1] Warsaw University of Technology, ul. Nowowiejska 15/19, 00-665 Warsaw, Poland
{Marek.Kozlowski,Karolina.Gabor-Siatkowska.dokt}@pw.edu.pl
{Marcin.Sowanski.dokt,Artur.Janicki}@pw.edu.pl
[2] Institute of Psychiatry and Neurology, ul. Sobieskiego 9, 02-957 Warsaw, Poland
istefaniak@ipin.edu.pl

Abstract. The article presents the results of work on improving sentiment and emotion recognition for Polish texts using a big data-based expansion process and larger neural language models. The proposed recognition method is intended to serve in a therapeutic dialogue system to analyze sentiment and emotion in human utterances. First, the language model is enhanced, by replacing the BERT neural language model with RoBERTa. Next, the emotion-based text corpus is enlarged. A novel process of augmenting an emotion-labeled text corpus using semantically similar data from an unlabeled corpus, inspired by semi-supervised learning methods, is proposed. The process of using the Common Crawl web archive to create an enlarged corpus, named CORTEX+pCC, is presented. An empathetic dialogue system named Terabot, incorporating the elaborated method, is also described. The system is designed to employ elements of cognitive-behavioral therapy for psychiatric patients. The improved language model trained on the enlarged CORTEX+pCC corpus resulted in remarkably improved sentiment and emotion recognition. The average accuracy and F1 scores increased by around 3% and 8% relative, which will allow the dialogue system to operate more appropriately for the emotional state of the patient.

Keywords: Natural language processing · Artificial intelligence · Big data · Humanities · Text-based emotion recognition · Corpus augmentation · Healthcare · Psychiatric therapy.

1 Introduction

Chatbots and dialogue systems are conversational agents that interact with users on a specific topic using natural language sentences. They are designed to interact

* This 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.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_33

verbally with people by analyzing spoken or written data, so that they can provide an appropriate response. Their popularity is rising in many fields, e.g., education and healthcare. Dialogue systems are created to perform some kind of action, therefore they will try to narrow down the conversation to get all needed information to do so. Chatbots, in contrast, are designed for extended conversations.

Goal-oriented dialogue systems, also called conversational, aim at completing a specific task through conversation with a user. Typically, such a system consists of a few sub-systems: natural language understanding with intent recognition and slot filling [5], dialogue state tracking [35], dialogue management [27], and language generation [24].

Recent development in conversational AI has been led almost exclusively by transformer-based neural language models specialized in dialogue generation. Since the initial release of a Transformer [30], or more specifically the Bidirectional Encoder Representations from Transformers (BERT) model [10], improvements to dialogue generation have been made mostly by increasing the number of network parameters. This strategy has proven to be so effective that in just a couple of years the set of problems in state-of-the-art models have shifted from problems with short, dull and uncontrollable answers to problems of how much understanding of meaning these models have. The most notable example of a conversational AI state-of-the-art model is ChatGPT, which uses GPT-4 [2] model, that is capable of attaining the highest possible score on AP Biology, AP Psychology, and a high score on Medical Knowledge Self-Assessment Program. While some researchers raise concerns about the reliability of the model's responses and its tendency to generate unrealistic content, others argue that ChatGPT has shown promise in the field of psychological therapy.

In recent years, there has been increasing acknowledgment of the important role mental health plays in achieving personal and global goals. The growing number of people with mental disorders forces one to look for new solutions to help patients [33]. One of the ideas is to support or supplement the therapist's work with tools that use new technologies. Some forms of therapy, e.g., cognitive behavioral therapy (CBT) or therapeutic techniques such as exposure in virtual reality settings, can, due to their well-described and structured nature, be applied in tools that use new technologies [11].

Chatbots have been tried for psychological therapies since the ELIZA chatbot [29]. A Woebot chatbot, working for the English language, has turned out to be helpful in therapy for depression [13]. Avatar-based dialogue systems (although human-controlled) have been shown to be successful in therapy for auditory hallucinations [6, 26]. The combined use of computer-based therapies may offer many possibilities for the treatment of physical and mental disabilities [12].

The work described in this article is part of the project devoted to the creation of a therapeutic dialogue system for Polish which would be able to realize elements of CBT by conducting empathetic dialogues with a patient. Such a system can support a human therapist, and free up some of a clinician's time; they can then use their time and skills, for example, to take better care of other

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_33

patients [4]. In this work we will propose such a dialogue system, named Terabot, and present its main components.

A dialogue system to be used in therapy needs not only to be able to respond according to the user's intent and the topics mentioned, but also to be empathetic. This means that its utterances should correspond with the user's emotional state. For this purpose, text-based sentiment and emotion recognition needs to be employed. In our previous work [36] we created such a module and tested it on a sentiment and emotion-labeled dataset, named CORTEX. In this work we propose big data- and deep learning-based methods to expand our training dataset. We show how they influence the sentiment and emotion recognition models.

2 Terabot – a therapeutic dialogue system

We designed a dialogue system, called Terabot, which will help in CBT therapy for psychiatric patients. It is equipped with a voice interface and is able to operate in Polish. Its main purpose is to help patients cope with difficult and overwhelming emotions influencing their life. In our study, we proposed the use of a dialogue agent, whose task will be to conduct exercises helping patients to cope with emotions (anxiety, anger, shame). The tool can be used as a supplement to therapies for many mental disorders, including psychotic disorders and personality problems.

We decided not to choose a neural dialogue system, due to the lack of sufficient training dialogue data. In addition, the lack of controllability of neural systems has been recently widely discussed [1, 32, 28]. Therefore, due to a very sensitive application context (therapy for a psychiatric patient), we decided to use a goal-oriented system, to retain better control of the system's actions, and to minimize the risk of giving an inappropriate answer to the patient.

A schematic diagram of Terabot is depicted in Figure 1. First, when a patient talks, the speech is recognized by the Google Web Speech API and transformed into text. Next, the text is further analyzed using the DIET Classifier [3] to identify the intentions and slots, and the sentiment/emotion recognizer, which is the part we focus on in this article. The DIET Classifier is pre-trained with the training data, originating from mock therapeutic sessions. Typical intents recognized by our dialogue system are: *Choose an exercise type*, *Describe the feeling*, *Explain the cause of the anxiety*. The slots are filled in with, e.g., an exercise name or recognized patient emotional state.

In the next step a decision is made about an action, resulting from a weighted combination of a rule policy, memoization policy (i.e., based on stories kept in memory) and Transformer Embedding Dialogue (TED) policy [31]. The TED policy takes into account the current dialogue state, i.e., among others, the patient intent(s) and the slot values. The latter include the current emotional state of the patient, recognized by the text-based sentiment/emotion recognition module, as described in this study.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_33

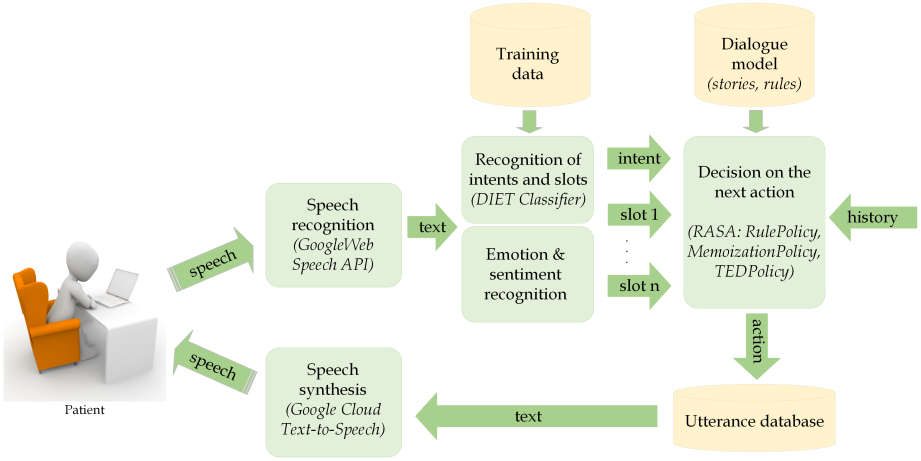

Fig. 1. Block diagram of the proposed therapeutic dialogue system.

If the next system action is an utterance (which is often the case), an appropriate text for the Terabot dialogue turn is found in the utterance database. Then, it is then transformed back to a speech signal by the Google Cloud Text-to-Speech (TTS) service. Both the DIET Classifier and the action decision pipeline are realized using RASA [14], an open-source framework for natural language understanding, dialogue management, and integration[3].

2.1 Sentiment and emotion classification

When designing Terabot, we evaluated several approaches to emotion and sentiment recognition, both simpler and more complex ones [36]. We started with baseline models such as multinomial Naïve Bayes and linear support vector machine (SVM) classifiers trained on top of bag-of-words (BoW) representation applied to token bigrams. Another model was based on the fastText algorithm, obtained from pre-trained word embeddings (300-dimensional variant) available in the *fasttext* library, for both English and Polish, and also using token bigrams.

The most complex approach was to fine-tune the pre-trained $\text{BERT}_{\text{BASE}}$ models. These models outperformed the simpler methods, reaching around 92% accuracy for sentiment classification and around 75% for emotion classification, for Polish datasets. The F1-score yielded similar values.

3 Datasets

3.1 CORTEX dataset

When working on the sentiment and emotion classifier for a Polish dialogue system, we faced the problem of lack of a relevant corpus for Polish which would

[3] https://rasa.com/products/rasa-platform/

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_33

contain conversational texts. Most of the existing dialogue corpora are either domain-specific and task-oriented or social media-oriented, and, therefore, they are difficult to apply in a therapeutic setting. There are, however, examples of emotionally grounded conversational datasets for English, such as DailyDialog [16] and EmpatheticDialogues [21], which are labeled with emotions at the utterance and dialogue levels, respectively.

The DailyDialog dataset (DD) consists of daily conversations obtained through crawling websites for learners of English. It counts about 13k dialogues, manually labeled with 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.

EmphatheticDialogues (ED) is a novel dataset with about 25k personal dialogues. Each dialogue is grounded in a specific situation where the speaker felt a given emotion, with the listener responding actively. This dataset is larger compared to DD, and experiments show that models built on this dataset are usually perceived by human evaluators as more empathetic, compared to models merely trained on large-scale Internet-crawled opinion-oriented data.

To create an emotion-labeled corpus with conversational data for Polish, we decided to use available resources for English and obtain the desired Polish version via neural machine translation (NMT). Since ED lacked examples of labeled neutral sentences, we added samples of neutral texts from the DD set. In this way, we built a parallel bilingual (English and Polish) corpus of emotional texts, labeled with three sentiment polarity classes and nine emotional classes. It is designed to serve experiments on sentiment and emotion recognition in a conversational context (e.g., dialogue systems or virtual assistants). We named the corpus CORTEX (for CORpus of Translated Emotional teXts) and made it publicly available[4].

During various experiments described in [36] we realized that one of the challenges is a limited number of examples in the training dataset (around 21k instances in total). One of the ways to augment the training dataset was to employ human annotators who would manually find and label new data. We did not opt for this method, as it is tedious, time-consuming, costly, and error-prone. We decided to invest in an automatic approach, which consists in identifying semantically similar sentences in big data: a web archive corpus, Common Crawl, described in the next section.

3.2 Common Crawl dataset

The exponential growth of the Internet community has resulted in the production of a vast amount of unstructured data, including web pages, blogs, and social media. Such a volume consisting of hundreds of billions of words is unlikely to be analyzed by humans. In this work we applied the *LanguageCrawl* tool [23], which

[4] CORTEX is freely available for download at https://github.com/azygadlo/CORTEX.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_33

allows natural language processing (NLP) researchers to easily build a single-language (e.g., Polish) web-scale corpus using the Common Crawl Archive [18, 25] – an open repository of web crawl information containing petabytes of data.

The Common Crawl Archive is built by the non-profit Common Crawl organization founded by Gil Elbaz in California [7]. Its purpose is to enable wider access to web information by manufacturing and to support an open web crawl repository. The data is made available both in raw HTML format (WARC) and text-only format (WET), and is freely accessible to anyone either via free direct download[5] or the commercial Amazon S3 service, for which a deposit is required.

The LanguageCrawl toolkit provides a highly concurrent actor-based architecture for building a local Common Crawl Archive. In [23] the authors presented three use cases: filtering of Polish websites, the construction of n-gram corpora, and the training of a continuous skipgram language model with hierarchical softmax. To create a Polish subset of Common Crawl, called hereinafter the Polish Common Crawl (pCC), the toolkit was applied in the first scenario (filtering websites).

Processing data from the Common Crawl Archive was a colossal task, which was severely limited by the Internet connection bandwidth, the cluster of multicore servers, and a bottleneck for data fetching. Therefore, it took several months to download enough data to construct a reasonable Polish website corpus.

Originally, the crawl data were stored in the Web Archive (WARC) format[6]. The WARC format retains and processes data from the Common Crawl Archive dump, which can be as large as hundreds of terabytes in size and contains billions of websites, in a more effective and manageable way. The raw crawl data is wrapped around the WARC format, ensuring a straightforward mapping to the crawl action. The HTTP request and response are stored, along with metadata information. In the case of the HTTP response, the HTTP header information is stored. This allows a high number of inquisitive insights to be gathered. The website content collected takes the form of an HTML document.

In our case, this rich format WARC is too complex, therefore we decided to use the WET Response format[7], which gathers plain-text web content instead of HTML. As most NLP tasks require only textual data, and we have access to limited resources, we decided to construct our tool around WET files, containing a minimal amount of metadata.

During our CORTEX expansion process we worked on corpora extracted from a few months of the 2015, 2018 and 2019 Common Crawl Archive, approximately 1PB in size and containing approximately 1.5 trillion web pages. Since we considered only WET files, we fetched around 100TB of compressed textual data (see the process sketched in Figure 2).Although the amount of data processed is enormous, the Polish language constitutes only a tiny fraction of it: we estimated it to be approximately 0.2 to 0.3%. Finally, our pipeline retrieved a few billion web pages with some Polish content. We decided to use only those web pages

[5] http://commoncrawl.org/
[6] https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml
[7] https://commoncrawl.org/the-data/get-started

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_33

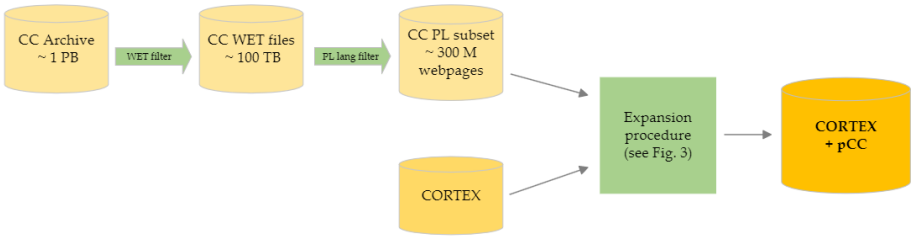

Fig. 2. Process of expanding the CORTEX database from big data perspective.

with most of the text written in Polish. This constraint led to 300 million web pages, which we used ultimately for the CORTEX expansion process. The whole filtering process, which employed 15 servers with 48 vCPUs and 256 GiB RAM each, took six months.

4 Experiments

4.1 Improving the classification model

When comparing various text-based emotion classification models in [36], the BERT models always yielded the best results. Since text representations generated by BERT have proved to be effective in many diverse NLP problems, e.g., classification, regression, machine translation, question answering, we continued to research these transformer-based language models in this project.

In the past three years, many modifications in the architecture of Transformer neural networks or their training procedures have been proposed. Transformers have the potential to learn longer dependencies, but are limited by a fixed-length context in the setting of language modeling. In [9] the authors proposed a novel neural architecture, Transformer-XL, which enabled learning dependency beyond a fixed length without disrupting temporal coherence. The authors of a Reformer model [15] introduced two techniques to improve the memory and training-time efficiency of large transformers, i.e., replacement of dot-product attention by one that uses locality-sensitive hashing, and applying reversible residual layers instead of the standard residuals, which allows the storing of activations only once in the training process. The XLNet model, presented in [34], was a generalized autoregressive pretraining method that enabled bidirectional context learning by maximizing the expected likelihood over all permutations of the factorization order. In addition, it overcame the limitations of the original BERT training procedure. The authors in [17] proved that BERT was significantly undertrained, and they proposed an optimized and larger neural model called Robustly Optimized BERT (RoBERTa).

In our work we decided to choose and evaluate a Polish version of RoBERTa. Although some other transformer-based language models for Polish are available, none has come even close to the scale, in terms of corpus size and number of

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_33

parameters, of the largest English-language models. In [8] two language models for Polish based on the RoBERTa approach were proposed. The larger model was trained on a dataset consisting of over 1 billion Polish sentences, or 135GB of raw text. We applied this model, called $\text{RoBERTa}_{\text{LARGE}}$, in the classification tasks, as an alternative to $\text{BERT}_{\text{BASE}}$.

The $\text{RoBERTa}_{\text{LARGE}}$ model has twice as many encoder blocks, more attention heads, and a richer token representation than $\text{BERT}_{\text{BASE}}$. As a result, it contains almost 3.5 times more parameters: 355M compared to 110M for $\text{BERT}_{\text{BASE}}$. We trained both models for 4 epochs with an effective batch size of 24, and, based on the validation metrics obtained after each training epoch, we selected the best models. The results will be presented and discussed in Section 5.

4.2 Dataset expansion process

In addition to improving the language model, we decided to extend the dataset used for training. Our idea was to find sentences in the pCC plain text corpora which would be semantically similar to sentences in the training part of the CORTEX dataset. Therefore, we sampled pCC to retrieve around 200 million web documents containing at least 10 sentences in a continuous manner. These sentences were considered for the potential expansion of CORTEX.

To resolve the problem of semantic similarity between sentences we used the state-of-the art sentence embeddings framework/library called SentenceTransformers[8] [22]. It is a Python framework for state-of-the-art sentence, text, and image embeddings. It is based on PyTorch and Transformers, and offers a large collection of pre-trained models tuned for various tasks. The framework can be used to compute sentence or text embeddings for more than 100 languages. These embeddings can then be compared, e.g., with cosine similarity to find sentences with a similar meaning.

Transformers were introduced in NLP in the late 2010s, to allow parallel computation and also to reduce the problem of long dependencies. The main advantages of transformers are: a) sentences are processed as a whole rather than word by word; b) self attention: to compute similarity scores between words in a sentence; c) positional embeddings: to encode information about position in the sentence.

The most popular NLP deep learning architecture, BERT, discussed previously in Section 4.1, is also based on transformers. BERT established a new state-of-the-art performance in many natural language understanding (NLU) tasks, e.g., in sentence-pair regression tasks like semantic textual similarity (STS). However, it requires that both sentences are fed into the network, which causes a massive computational overhead: finding the most similar pair in a collection of 10,000 sentences requires about 50 million inference computations. The construction of BERT makes it unsuitable for either semantic similarity search or for unsupervised tasks like clustering.

[8] https://www.sbert.net

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_33

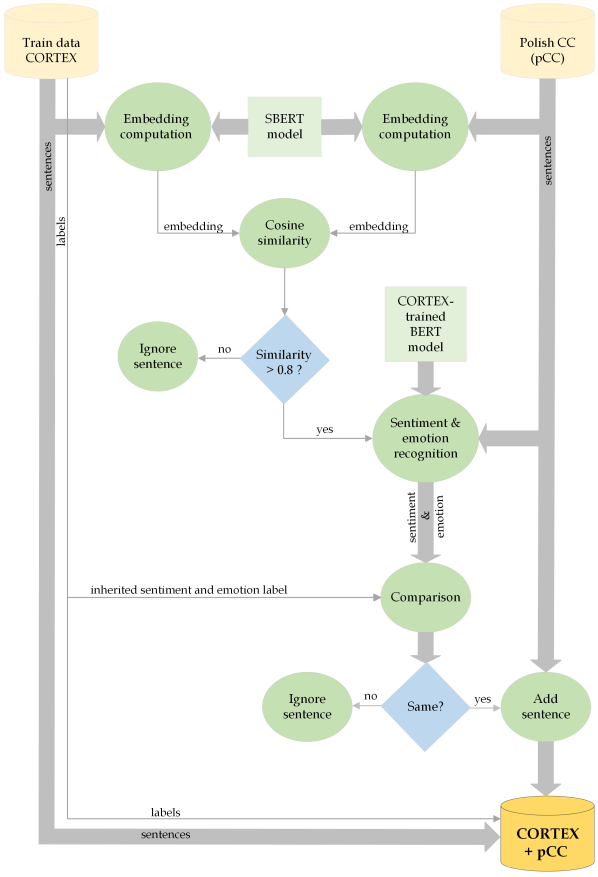

Fig. 3. Scheme of the formation of the improved CORTEX database with the Polish Common Crawl.

In [22], the authors proposed Sentence-BERT (SBERT), a modification of the pre-trained BERT network that uses Siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine similarity. This reduces the effort of finding the most similar pair from 65 hours with BERT/RoBERTa to about 5 seconds with SBERT, while maintaining the accuracy of BERT.

Therefore, we decided to use the SBERT-based approach to compute embeddings of sentences coming from the CORTEX training dataset and pCC, and next we evaluated their semantic similarity using cosine measure between embeddings. More precisely, for each sentence/prompt from the CORTEX training subset we found the top-10 semantically closest ones within pCC. In addition, we decided that their cosine similarity must be over a threshold of 0.8 (the value was set heuristically). This way we expanded the CORTEX prompts with se-

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_33

Table 1. Train dataset statistics for sentiment and emotion labels schemas, for CORTEX (original) and CORTEX expanded with CommonCrawl web sites (CORTEX+pCC).

Sentiment recognition schema			Emotion recognition schema		
Sentiment	CORTEX	CORTEX+pCC	Emotion	CORTEX	CORTEX+pCC
Positive	5985	24608	Happiness	2391	11107
			Confidence	1705	6460
			Other positive	1889	7041
Negative	8314	26768	Sadness	2267	8093
			Anger	1804	5243
			Fear	1565	5759
			Guilt	1576	5018
			Other negative	1102	2655
Neutral	7149	28157	Neutral	1787	28157

mantically similar ones from pCC, and assigned them corresponding categories (mood, sentiment) of the given reference CORTEX prompt. In other words, for a given CORTEX prompt (*query*) we found the pCC-origin semantically closest sentences (*candidates*) and assigned *candidates* the same categories (mood, sentiment) as were in the *query*.

Next, we analyzed the expanded sentences with the assigned CORTEX categories using the original BERT classifiers trained on the original CORTEX training dataset. We predicted categories (mood, sentiment) for each "expanded" sentence. If they were the same as those assigned during the first expansion phase, we retained the given sentence in the extended dataset. If not, such a sentence was removed. The proposed approach resembles the semi-supervised learning (SSL) technique [19, 20], in which a large unannotated dataset is assigned labels based on a classifier trained on a much smaller annotated dataset.

The whole process of CORTEX corpus expansion is shown in Figure 3. It took four weeks on a single server with four V100 GPUs and 32 GiB RAM memory each. As a result, we created a new extended training dataset, which is almost four times larger than the original CORTEX training dataset, which contained 21,800 records (sentences). After expansion, the training set reports around 79,500 records. This dataset will hereinafter be referred to as CORTEX+pCC. The detailed statistics of the training set are shown in Table 1.

5 Results and Discussion

To verify the impact of the proposed changes on sentiment and emotion recognition, we compared it against the baseline results presented in [36]. We used the same test set as in [36] and observed precision, recall, F1-score, accuracy (micro F1), macro averaged F1, and support-weighted F1. First, we will present the results achieved by using an improved language model. Next, in addition to the model change, we will show the impact of using an extended training dataset.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_33

5.1 Impact of improved language model

Table 2 shows the comparison of text-based sentiment and emotion recognition for Polish, using the $\text{BERT}_{\text{BASE}}$ and $\text{RoBERTA}_{\text{LARGE}}$ neural language models. The training and testing datasets remained unchanged.

Unsurprisingly, for the 3-class scenario, evaluation metrics reached much higher values than for the 9-class scenario, because of the number of degrees of freedom. We reached almost 94% accuracy for sentiment classification and over 80% for emotion classification. The F1-score yielded similar values.

Table 2. Sentiment and emotion classification results for Polish $\text{BERT}_{\text{BASE}}$ and $\text{RoBERTa}_{\text{LARGE}}$ models trained on original CORTEX trainset.

Task	Metric	$\text{BERT}_{\text{BASE}}$	$\text{RoBERTa}_{\text{LARGE}}$
Sentiment (3 class)	Accuracy	0.9224	0.9385
	Weighted F1	0.9226	0.9385
Emotion (9 class)	Accuracy	0.7519	0.8040
	Weighted F1	0.7515	0.8035

The results confirmed that $\text{RoBERTA}_{\text{LARGE}}$ yielded better results than $\text{BERT}_{\text{BASE}}$: the accuracy and F1 of sentiment recognition for Polish increased by more than 1.5% relative. As for emotion recognition, the advantage was even clearer: the accuracy and F1-score (weighted) increased by more than 5% relative.

We found this improvement remarkable, but expected there was still room for improvement using the training dataset expansion. These results are discussed in the next subsection.

5.2 Impact of extended training dataset

We decided to train the $\text{RoBERTA}_{\text{LARGE}}$ model, which yielded better results in the previous experiment, with the extended training corpus CORTEX+pCC. Next, we tested it using the same test set as in the previous experiment.

We evaluated the results of sentiment and emotion recognition; the results are displayed in Tables 3 and 4, respectively. We conducted many analyses including quality assessment per category, and aggregated measures (micro, macro, weighted), comparing the baseline results ($\text{BERT}_{\text{BASE}}$ trained on the original CORTEX corpus) with the proposed approach, i.e., the $\text{RoBERTA}_{\text{LARGE}}$ model trained on the extended CORTEX+pCC training dataset.

We observed a significant quality improvement in favor of the proposed approach. We outperformed the original approach in both tasks: sentiment classification and emotion classification. In sentiment classification we reported accuracy (micro F1) and weighted-averaged F1 around 95%, i.e., 3% relative higher than in the original approach. The best predictions concerned the neutral label (F1 = 98%, slight improvement from 97% in the original approach). The best partial F1 improvement was reported for positive label, from 88% to 92%.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_33

Table 3. Results of sentiment classification for frozen test data for BERT/RoBERTa model trained on two Polish datasets: CORTEX (original) and CORTEX expanded with CommonCrawl web sites (CORTEX+pCC).

Sentiment	Support	Metric	BERT/ CORTEX	RoBERTa/ CORTEX+pCC
Positive	826	Precision	0.88	0.93
		Recall	0.89	0.90
		F1-score	0.88	0.92
Negative	1084	Precision	0.91	0.93
		Recall	0.92	0.95
		F1-score	0.91	0.94
Neutral	955	Precision	0.98	0.98
		Recall	0.95	0.97
		F1-score	0.97	0.98
Average	2865	Accuracy	0.92	0.95
		Macro avg.	0.92	0.94
		Weighted avg.	0.92	0.95

We also achieved remarkable gains in the emotion-classification problem, which is a more complex problem because of the number of classes. During emotion-detection evaluation we noticed that all aggregated quality measures, i.e., accuracy, macro-averaged F1, and weighted-averaged F1, are about 8% relative higher than the baseline measures.

The three best predictive labels were neutral (F1 = 96%), fear, and sadness (F1 = 84%). The best partial F1 improvement was reported for labels confidence, sadness, and guilt; the quality gains in those categories were around 8-9% relative, in comparison with the previously scored values presented in [36]. In terms of the application of the emotion recognizer in the empathetic dialogue system, such an improvement in detecting such sensitive emotions was highly desired.

6 Conclusions

In this article we presented a novel intelligent process of enlarging (augmenting) an emotion-labeled text corpus using semantically similar sentences from big data: a massive yet unlabeled corpus – in this case the Common Crawl web archive. The proposed approach was inspired by the SSL approach, known in other areas of machine learning. Similarly to SSL applications, here we also observed that using an *imperfect* sentiment and emotion classifier (taken from [36]) we were able to automatically enlarge the training dataset using unlabeled data, which ultimately has led to constructing a *strongly improved* classifier.

We also proposed replacing the previously used $\text{BERT}_{\text{BASE}}$ neural language model with a model with richer architecture: $\text{RoBERTa}_{\text{LARGE}}$. This model, trained with the enlarged CORTEX+pCC corpus, allowed for an improvement

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_33

Table 4. Results of emotion classification with BERT/RoBERTa model trained on original trainset (CORTEX) and data expanded with CommonCrawl web sites (CORTEX+pCC).

Emotion	Support	Metric	BERT/ CORTEX	RoBERTa/ CORTEX+pCC
Anger	226		0.70	0.77
Confidence	222		0.75	0.83
Fear	208		0.80	0.84
Guilt	199		0.73	0.82
Happiness	328	F1-score	0.75	0.83
Neutral	238		0.92	0.96
Other negative	160		0.66	0.79
Other positive	276		0.67	0.76
Sadness	291		0.76	0.84
		Accuracy	0.75	0.83
Average	2148	Macro avg.	0.75	0.83
		Weighted avg.	0.75	0.83

of text-based sentiment and emotion recognition for Polish: the average accuracy and F1 increased by around 3% and 8%, respectively. The ultimate accuracy scores reached 95% and 83% for sentiment and emotion recognition, respectively. However, despite such progress we must be aware that the recognition is still not error-free.

Last but not least, in this article we presented an empathetic dialogue system for therapeutic purposes, named Terabot, for which the proposed text-based sentiment and emotion-recognition module is intended. Thanks to the remarkably improved recognition of guilt, happiness, sadness, and other emotions, the dialogue system will work better to understand the emotional state of a patient undergoing therapy.

References

1. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., et al.: On the opportunities and risks of foundation models. CoRR (2021). https://doi.org/10.48550/arXiv.2108.07258, https://arxiv.org/abs/2108.07258
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
3. Bunk, T., Varshneya, D., Vlasov, V., Nichol, A.: Diet: Lightweight language understanding for dialogue systems (2020). https://doi.org/10.48550/arXiv.2004.09936, https://arxiv.org/abs/2004.09936
4. Carroll, K., Rounsaville, B.: Computer-assisted therapy in psychiatry: Be brave—it's a new world. Current psychiatry reports **12**, 426–32 (10 2010). https://doi.org/10.1007/s11920-010-0146-2
5. Chen, Q., Zhuo, Z., Wang, W.: BERT for joint intent classification and slot filling. arXiv preprint arXiv:1902.10909 (2019)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_33

6. Craig, T.K., Rus-Calafell, M., Ward, T., Leff, J.P., Huckvale, M., Howarth, E., Emsley, R., Garety, P.A.: Avatar therapy for auditory verbal hallucinations in people with psychosis: a single-blind, randomised controlled trial. Lancet Psychiatry **5(1)**, 31–40 (2018). https://doi.org/10.1016/S2215-0366(17)30427-3
7. Crouse, S., Nagel, S., Elbaz, G., Malamud, C.: Common Crawl Foundation. http://commoncrawl.org (2008)
8. Dadas, S., Perełkiewicz, M., Poświata, R.: Pre-training polish transformer-based language models at scale. In: Artificial Intelligence and Soft Computing. pp. 301–314. Springer International Publishing (2020)
9. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.V., Salakhutdinov, R.: Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860 (2019)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
11. Dino, F., Zandie, R., Abdollahi, H., Schoeder, S., Mahoor, M.H.: Delivering cognitive behavioral therapy using a conversational social robot. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2089–2095 (2019). https://doi.org/10.1109/IROS40897.2019.8968576
12. Fernández-Caballero, A., Navarro, E., Fernández-Sotos, P., González, P., Ricarte, J., Latorre, J., Rodriguez-Jimenez, R.: Human-avatar symbiosis for the treatment of auditory verbal hallucinations in schizophrenia through virtual/augmented reality and brain-computer interfaces. Frontiers in Neuroinformatics **11** (11 2017). https://doi.org/10.3389/fninf.2017.00064
13. 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 **4**(2), e19 (Jun 2017). https://doi.org/10.2196/mental.7785, http://mental.jmir.org/2017/2/e19/
14. Jiao, A.: An intelligent chatbot system based on entity extraction using RASA NLU and neural network. Journal of Physics: Conference Series **1487**(1), 012014 (mar 2020). https://doi.org/10.1088/1742-6596/1487/1/012014
15. Kitaev, N., Kaiser, Ł., Levskaya, A.: Reformer: The efficient transformer. arXiv preprint arXiv:2001.04451 (2020)
16. 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). pp. 986–995. Asian Federation of Natural Language Processing, Taipei, Taiwan (Nov 2017), https://www.aclweb.org/anthology/I17-1099
17. 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)
18. Mühleisen, H., Bizer, C.: Web data commons-extracting structured data from two large web corpora. LDOW **937**, 133–145 (2012)
19. Oliver, A., Odena, A., Raffel, C.A., Cubuk, E.D., Goodfellow, I.: Realistic evaluation of deep semi-supervised learning algorithms. Advances in Neural Information Processing Systems **31** (2018)
20. Pudo, M., Szczepanek, N., Lukasiak, B., Janicki, A.: Semi-supervised learning with limited data for automatic speech recognition. In: Proc. IEEE 7th Forum on Research and Technologies for Society and Industry Innovation (RTSI 2022). pp. 136–141. Paris, France (2022). https://doi.org/10.1109/RTSI55261.2022.9905112

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_33

21. Rashkin, H., Smith, E.M., Li, M., Boureau, Y.L.: Towards empathetic open-domain conversation models: A new benchmark and dataset. arXiv preprint arXiv:1811.00207 (2018)
22. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. arXiv preprint arXiv:1908.10084 (2019)
23. Roziewski, S., Kozłowski, M.: LanguageCrawl: A generic tool for building language models upon Common Crawl. Language Resources and Evaluation **55**(4), 1047–1075 (2021)
24. Sharma, S., He, J., Suleman, K., Schulz, H., Bachman, P.: Natural language generation in dialogue using lexicalized and delexicalized data. In: International Conference on Learning Representations: Workshop (2017)
25. Smith, J.R., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., Lopez, A.: Dirt cheap web-scale parallel text from the common crawl. In: ACL (1). pp. 1374–1383 (2013)
26. Stefaniak, I., Sorokosz, K., Janicki, A., Wciórka, J.: Therapy based on avatar-therapist synergy for patients with chronic auditory hallucinations: A pilot study. Schizophrenia Research **211**, 115–117 (2019). https://doi.org/https://doi.org/10.1016/j.schres.2019.05.036, https://www.sciencedirect.com/science/article/pii/S0920996419302130
27. Su, P.H., Gasic, M., Mrkšić, N., Barahona, L.M.R., Ultes, S., Vandyke, D., Wen, T.H., Young, S.: On-line active reward learning for policy optimisation in spoken dialogue systems. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2431–2441 (2016)
28. Tamkin, A., Brundage, M., Clark, J., Ganguli, D.: Understanding the capabilities, limitations, and societal impact of large language models. CoRR (2021). https://doi.org/10.48550/arXiv.2102.02503, https://arxiv.org/abs/2102.02503
29. Vaidyam, A.N., Wisniewski, H., Halamka, J.D., Kashavan, M.S., Torous, J.B.: Chatbots and conversational agents in mental health: a review of the psychiatric landscape. The Canadian Journal of Psychiatry **64**(7), 456–464 (2019)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
31. Vlasov, V., Mosig, J.E.M., Nichol, A.: Dialogue transformers (2019). https://doi.org/10.48550/arXiv.1910.00486, https://arxiv.org/abs/1910.00486
32. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al.: Ethical and social risks of harm from language models. CoRR (2021). https://doi.org/10.48550/arXiv.2112.04359, https://arxiv.org/abs/2112.04359
33. World Health Organization, et al.: The WHO special initiative for mental health (2019-2023): universal health coverage for mental health. Tech. rep., World Health Organization (2019)
34. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems **32** (2019)
35. Zhong, V., Xiong, C., Socher, R.: Global-locally self-attentive dialogue state tracker. In Association for Computational Linguistics (2018)
36. Zygadło, A., Kozłowski, M., Janicki, A.: Text-based emotion recognition in English and Polish for therapeutic chatbot. Applied Sciences **11**(21), 10146 (2021)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_33