

CLARIN-Emo: Training Emotion Recognition Models using Human Annotation and ChatGPT*

Bartłomiej Koptyra^[0009-0005-9938-305X], Anh Ngo^[0009-0002-6902-007X],
Łukasz Radliński^[0000-0002-7366-3847], and Jan Kocoń^[0000-0002-7665-6896]

Department of Artificial Intelligence
Wrocław University of Science and Technology, Wrocław, Poland
bartlomiej.koptyra@pwr.edu.pl

Abstract. In this paper, we investigate whether it is possible to automatically annotate texts with ChatGPT or generate both artificial texts and annotations for them. We prepared three collections of texts annotated with emotions at the level of sentences and/or whole documents. CLARIN-Emo contains the opinions of real people, manually annotated by six linguists. Stockbrief-GPT consists of real human articles annotated by ChatGPT. ChatGPT-Emo is an artificial corpus created and annotated entirely by ChatGPT. We present an analysis of these corpora and the results of Transformer-based methods fine-tuned on these data. The results show that manual annotation can provide better-quality data, especially in building personalized models.

Keywords: ChatGPT · Emotion recognition · Automatic annotation.

1 Introduction

Emotions play an important role in human lives. Expressing and detecting them is a vital skill in social interactions, including professional careers. We even try to measure our abilities in this aspect, calling them emotional intelligence. Therefore it is no wonder that many studies concern themselves with trying to detect and use publicly expressed emotions to improve on a wide range of tasks such as stock market prediction [14], understanding user-preferences [28], public health monitoring and surveillance [27].

Understanding human emotions is one of the more challenging tasks in natural language processing. Not only are they a very subjective topic, but humans can also lack the capability to express themselves in written language. This inaccuracy of expressed emotions is because more natural, multimodal forms of

* This work was financed by the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme: (1) POIR.04.02.00-00C002/19 (AN,LR); (2) POIR.01.01.01-00-0615/21 (BK); (3) POIR.01.01.01-00-0288/22 (JK); (4) the statutory funds of the Department of Artificial Intelligence, Wrocław University of Science and Technology; (5) the European Union under the Horizon Europe, grant no. 101086321 (OMINO).

emotional expression, such as speech, body language, and facial expressions, can, unlike text, portray peculiar cues used to recognize emotions [1]. The difference between formal and informal writing styles, accounting for grammatical errors (especially in short texts) [1], use of figurative language like sarcasm [22], and many more are aspects that contribute to the difficulty. This emphasizes a need for good training data. This is not a trivial task, as creating a dataset with emotion labels requires a set of guidelines for annotators to follow for annotations to be consistent. Examining the publicly available datasets for emotion recognition in the Polish language reveals the lack thereof. CLARIN-Emo dataset was created as a subset of a Polish sentiment annotated consumer reviews corpus PolEmo [19, 18, 37, 15, 20, 3, 23] with additional emotion annotations to fill this gap.

With the recent rise in popularity of large language models, we took this opportunity to explore the possibility of creating consistent emotion annotations using such a model, namely chatGPT, on a different data set and compare it against a model trained and tested on the newly created human-annotated emotions dataset. This allows checking the consistency of such annotations by training a model on them. We also explore the possibility of not only annotating but also generating already annotated data using this multi-purpose dialog system created by OpenAI. Such a system requires dialogue breadth (coverage) and dialog depth (complexity). Emotion recognition and expression are important features that can increase dialog depth [5]. Checking the consistency of emotion recognition of such a system helps to recognize at what level of dialog depth are modern, popular large language models. Our findings suggest that such models are still not up to human standards.

We use three created datasets to train and evaluate multilabel emotion classification models, on the level of both documents and sentences, by fine-tuning pre-trained transformers. We treat sentences as a part of the document to train sequential sentence classification models. We also perform a personalized setup for sequence sentence classification in the new CLARIN-Emo dataset, similar to [17, 26, 13], and find that personalization has a positive impact on model performance in emotion recognition.

2 Related work

There are many non-Polish datasets for emotion recognition that follow different emotion taxonomies. The most basic approach is to use a distinction proposed by Ekman [11], which states six basic human emotions: anger, surprise, disgust, enjoyment, fear, and sadness. Such an approach can be found, for example, in the EmotionLines dataset [5], which assigns emotions to utterances in a dialog setting to achieve the contextual flow of emotions, and in ArmanEmo [25], which is a sentence-level human-labeled annotated dataset based on Twitter, Instagram, and comments from an Iranian e-commerce company. Plutchnik's "wheel of emotions" [30] extended this notion into four contrasting emotion pairs: joy vs sadness, anger vs fear, trust vs disgust, and surprise vs anticipation. It is a

popular approach used in many datasets because of its relatively small number of classes and the use of the wheel as representation. This representation allows using angles in the wheel as distances between emotions. Plutchnik’s set of emotions can be found used in creating a word-emotion association lexicon [4], even in crowdsourcing setting, assessing the psychology of story characters [31], emotion detection on social media [35], and many more. Plutchnik’s set of emotions can also be expanded into 24 emotions with inter-annotator agreement based on distances on the wheel [10]. There are also other datasets with more emotions, like 28, used in the GoEmotions dataset [9], which trade F1-score for finer granularity.

Text classification has been a prominent field of research for many years, with recent years bringing a shift towards more advanced deep learning algorithms [24]. The latter dominated the field over the last few years with particular transformer models such as RoBERTa [21], XML-RoBERTa [8] and DistilBERT [34] achieving impressive results for many text classification tasks.

When it comes to research on ChatGPT for text processing tasks, most of the studies conducted so far mainly focus on text summarisation [12, 29] and question answering [36, 2]. Very few explore such areas as sentiment analysis [38] or humor recognition and generation [6]. However, there is a large-scale evaluation of ChatGPT that has been recently published [16], which mainly focused on exploring ChatGPT performance when compared to SOTA solutions for many different text processing tasks, including text classification and sentiment analysis. ChatGPT didn’t outperform any SOTA solutions. However, it performed significantly worse on most emotion and sentiment processing tasks. There haven’t been any major studies conducted that would explore the potential usage of the chatGPT for data annotation or generation of synthetic datasets for text processing tasks.

3 Datasets

As part of our work, we have prepared several datasets that will be used to compare different approaches to emotion recognition. The first dataset (CLARIN-Emo) contains Polish reviews written by humans and manually annotated with emotions, also by humans. The second dataset (Stockbrief-GPT) contains Polish business articles written by humans and annotated with emotions using ChatGPT. The third dataset contains English reviews, where ChatGPT generated both the text and the emotion annotations.

3.1 CLARIN-Emo: Human Texts and Annotations

The CLARIN-Emo dataset contains a subset of consumer reviews from the PolEmo 2.0 corpus. The reviews belong to one of four domains: *hotels*, *medicine*, *products*, and *school*. The collection also includes non-opinion texts, which are informative, belong to the same domains, and are mostly neutral. The CLARIN-Emo corpus contains a total of 1110 opinions composed of 8891 sentences. Annotators labeled the sentiment and emotions of the opinions at the level of each

sentence and the whole opinion. The emotion labels are taken from Plutchik’s model. There are eight emotion dimensions: *joy*, *trust*, *anticipation*, *surprise*, *fear*, *sadness*, *disgust*, and *anger*. In addition, three sentiment dimensions were used: positive, negative, and neutral. The corpus was annotated by six people who did not see each other’s decisions. The annotation was multi-labeled – several labels could describe each sentence and each opinion. Table 1 shows the agreement of annotations for each emotion dimension and the total agreement for all dimensions. Those labels that were annotated at least twice by different annotators were included in the final corpus. The same table 1 also summarizes the number of annotations in the final corpus for each label, assigned at the sentence and text levels.

Table 1: Annotations statistics in CLARIN-Emo dataset.

Emotion	Number of instances		Krippendorff’s alpha coefficient	
	Document-level	Sentence-level	Document-level	Sentence-level
Joy	626	4222	0.6497	0.7190
Trust	274	1972	0.3514	0.3990
Anticipation	92	1078	0.1258	0.3307
Surprise	72	582	0.1118	0.3497
Fear	50	370	0.1214	0.3436
Sadness	604	3892	0.5145	0.6481
Disgust	313	1664	0.2359	0.4276
Anger	274	1406	0.4238	0.4877
Positive	665	4687	0.7116	0.7317
Negative	618	4122	0.7510	0.8195
Neutral	176	2364	0.5908	0.5990
Total	3764	26359	0.4171	0.5323

3.2 Stockbrief-GPT: Human Texts Annotated by ChatGPT

Stockbrief-GPT is a corpus of Polish articles on economics containing 227 articles and 3795 sentences. We used ChatGPT to annotate these articles at the sentence and whole-text levels. Below is a translation of the sample prompt and the system’s response.

Task: Stockbrief-GPT (example translated from Polish to English).

Prompt

Company: cd project

Rate the following sentences against the Company given earlier using the following label identifiers from the specified categories: "S:sentiment" (-3:strongly negative, -2:negative, -1:weakly negative, 0:neutral, 1:weakly positive, 2:positive, 3:strongly positive), "E:emotion" (fear, stress, sadness, anger, joy, trust, surprise, anticipation).

Additionally, evaluate the entire text using the same labels. Present the result as a JSON object containing the following fields: { "sentences": { "sentence_number": {"S": identifier_label_sentence, "E": list_label_emotion}}, "evaluation_text": { "S": identi-

```
fier_label_sentiment, "E": list_label_emotion}}.
```

1. The CD Project's poker play. **2.** On Monday, after the trading session, the world of listed game companies was flooded with information that all investors interested in domestic gaming had been waiting for - CD Projekt officially confirmed that it was working on the fourth installment of the "Witcher" game series. **3.** However, the developers have indicated that it will not be a game called "The Witcher 4", as the market had expected. **4.** As originally expected, the new installment of the Witcher saga was to be titled 'The Witcher 4' and tell the story of Ciri - Geralt's ward, wearing a medallion from the Witcher School of the Cat. (...)

ChatGPT answer

```
{ "sentences": { "1": {"S": 1, "E": ["anticipation"]},
"2": {"S": 2, "E": ["anticipation"]},
"3": {"S": -1, "E": ["surprise"]},
"4": {"S": 0, "E": ["anticipation"]}, ...}, "evaluation_text": {"S": 0, "E": [ "anticipation" ] }}
```

Table 2 how many occurrences of each label are in the Stockbrief-GPT dataset. As strong positive and negative labels were very sparse in the Stockbrief-GPT dataset for our experiments, we cast those labels as positive and negative, respectively.

Table 2: Number of annotations of different emotions in Stockbrief-GPT dataset.

Emotion	Document-level	Sentence-level
Fear	6	97
Stress	21	171
Sadness	47	415
Anger	35	347
Joy	75	895
Trust	44	403
Suprise	12	219
Anticipation	123	1247
Strong negative	0	32
Negative	18	481
Weak negative	45	415
Neutral	40	904
Weak positive	94	797
Positive	30	1135
Strong positive	0	31
Total	590	7589

3.3 ChatGPT-Emo: ChatGPT Texts and Annotations

ChatGPT-Emo dataset contains opinions and reviews with corresponding domains and emotional labels of the opinion. The initially generated dataset comprised 6684 texts with 3486 unique emotional labels and 613 domains. ChatGPT fully generated the dataset. Below is an example of the prompt and the system's answer:

Task: ChatGPT-Emo.	
Prompt	Create example dataset containing 3 long opinions (at least 5-10 sentences each) form different domains and labels described emotions evoked by these opinions. The output should be an array of JSON objects containing the following keys: "text" is a text of the opinion, "domain" is the domain of the opinion, and "labels" is an array of emotions evoked by the opinion. Do not explain yourself at all, ONLY array of JSONs should be you answer. Do not make textual introduction before the array. Start the texts differently, make them as real opinions of users from the web. Each opinion should start with different word. If you know real opinions from the web, use them. Use from 5 to 10 emotional labels for each example.
ChatGPT answer	<pre>[{"text": "I recently attended a concert by this band and it was an unforgettable experience. The energy and passion they brought to the stage was electrifying, and the crowd was completely enthralled. The music was amazing, and I was dancing and singing along the entire time. I would highly recommend this band to anyone looking for a high-energy and entertaining concert experience. The stage production was also top-notch, and I was blown away by the light show and special effects. I can't wait to see them again.", "domain": "Entertainment", "labels": ["Excitement", "Thrills", "Engagement", "Entertainment", "Admiration", "Satisfaction", "Anticipation", "Awe", "Enthrallment", "Unforgettable"]}, ...]</pre>

Many of the initially generated emotional labels were synonymical to each other e.g. *joy*, *joyful*, *happy*, *happiness* etc., thus, it was important to map the similar labels to each other. In order to achieve that, a three-step approach was used. Firstly, all labels with the same lemma recognized by the spaCy English language model were mapped to the lemma of the labels. Then lemmas with similar meaning were grouped manually into 8 major categories: *disappointed*, *angry*, *disgusted*, *excited*, *happy*, *satisfied*, *sad* and *rare* where rare included all labels that didn't have synonymous labels in the dataset and appeared 50 or less times in the entire dataset. Then all duplicate texts and all texts with only rare labels assigned were removed from the dataset. Domains were manually mapped into 13 categories: emphScience, Games & Technology, Art, Culture & Entertainment, Food & Dining, Health & Fitness, Travel, Cities & Tourism, Cars, Workplace & Economics, Home, Animals & Nature, Shopping & Products, Personal & Lifestyle, Politics & Social issues, Education. Final dataset consists of 5959 texts of 13 domains with 7 unique emotional labels assigned.

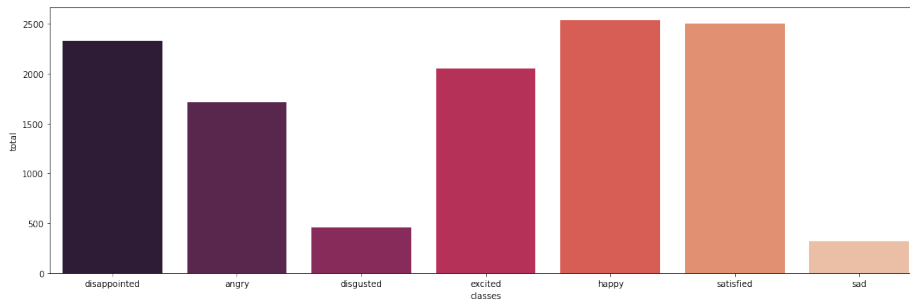


Fig. 1: ChatGPT-Emo distribution of texts into 7 classes: disappointed, angry, disgusted, excited, happy, satisfied, sad

Figure 1 shows the distribution of different classes. It can be observed that most 'positive labels' occur with similar frequency and the 'negative' labels have a greater variety of occurrence. An argument could be made since categories *satisfied* and *happy* have similar meanings and almost identical frequency of occurrence. However, the category *satisfied* was used to group labels that were positive but not explicitly emotional such as *fulfill* and *content*. *Happy* was used to more explicitly emotional labels such as *joy* and *pleased*.

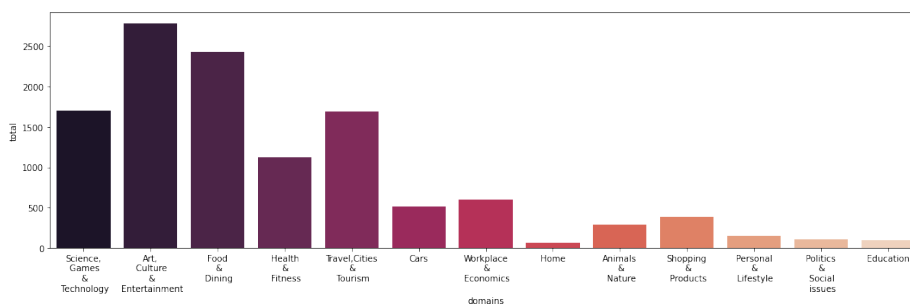


Fig. 2: ChatGPT-Emo distribution of texts into 13 domains: Science, Games & Technology, Art, Culture & Entertainment, Food & Dining, Health & Fitness, Travel, Cities & Tourism, Cars, Workplace & Economics, Home, Animals & Nature, Shopping & Products, Personal & Lifestyle, Politics & Social issues, Education

Figure 2 shows the distribution of texts into 13 distinguished domains. As seen in the figure most common categories are broad categories such as *Art, Culture & Entertainment* and *Food & Dining* and more narrow ones such as *Home* or *Education* are among the most rarely occurring ones. Surprisingly many reviews regarded *cars* even though it is arguably the most narrow category of all listed.

4 Models

4.1 Sequential Sentence Classification

For the Sequential Sentence Classification (SSC) task, given a sequence of sentences $X = x_1, x_2, \dots, x_N$ where N is the number of sentences, the goal is to predict the set of labels for each sentence based on the context in which the sentences appear. Our model is closely derived from the approach proposed by [7], in which the authors proposed a novel input representation and utilized the BERT model by fine-tuning the target SSC tasks to classify sentences in scientific abstracts into their rhetorical roles. This section describes the model along with our modifications.

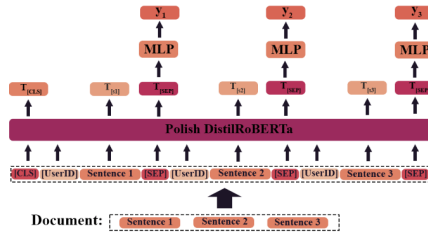


Fig. 3: Overview architecture of Sequential Sentence Classification

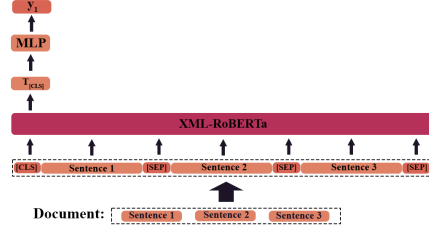


Fig. 4: Overview architecture of Text Classification

Figure 3 illustrates the model overview. It employed the two special tokens of BERT-based architecture, [CLS] and [SEP]. As input to BERT, a given sequence of sentences $X = x_1, x_2, \dots, x_N$ is concatenated by separator tokens [SEP], resulting in a whole long sequence, before adding the standard classification token [CLS] at the beginning. However, unlike the standard BERT, which uses the [CLS] token as the input representation of the entire sequence, the proposed method uses the [SEP] token encodings as the input representation fed into a multi-layer feedforward network for classification. The author’s idea behind using [SEP] to represent a sequence of sentences is that the [SEP] tokens learn sentence structure and the relations between consecutive sentences through BERT’s pre-training, allowing the BERT’s self-attention layer to leverage contextual information from all the words in the entire document, while still benefitting from BERT’s pre-trained weights.

Loss function. The original model is trained on multi-class classification tasks and aims to minimize cross-entropy loss between the predicted probability distribution and the target distribution of the classes. In addition, when we conducted some initial experiments, we noticed that the model struggled severely with unbalanced label distributions. Since each sentence in our dataset could have multiple labels, we introduced the positive weights to the BCEWithLogitLoss, which are calculated by weighing the samples as the inverse of the class frequency for its classes in the training set. The formula followed the previous study from [33]:

$$weight_{class_n} = \frac{\sum_{i=1}^n x_n}{x_n}, \quad (1)$$

where x_n is the count of samples of class n in the training set

Personalized approach. Since emotion recognition is a subjective task [17], each annotator could have a different perspective on the same sentence. We decided to combine the SSC model with a personalized approach called UserID proposed by [26]. In this method, each annotator is represented by a special token in text embedding, by which the model is provided with information about each annotator.

4.2 Text Classification

The main goal of the text classification task is to assign a set of labels to the given record of text. In the case of our paper, the main goal is to assign the set of emotional labels from the given set to each text in the given dataset. The model we use is the base version of XML-RoBERTa [8]. This version of the BERT model is trained in multiple languages, which is important in this case as two of the tested datasets are in Polish, and one is in English. As seen in Fig. 4 the architecture of the model is way simpler than for sequence classification, as there is no need to generate separate outputs for each sentence. Opposite to the first architecture in this case the [CLS] token is used as the input to MLP which then assigns the set of labels to the given text.

5 Experiments

5.1 Sequential Sentence Classification

We formulated experiments as multi-label classification on two datasets Clarin-Emo and Stockbrief-GPT, which aim to predict emotional labels for each sentence in a given document. The macro F1-score and micro F1-score were used for model evaluation. Both datasets were split into Train (70%), Dev (15%), and Test (15%). Except for the experiment with the personalized method UserID where the labels from six annotators are considered individually, in other experiments concerning CLARIN-Emo, these labels are aggregated using majority voting.

Language models. All experiments utilized the Polish DistilRoBERTa multilingual pretrained model proposed by [32]. We also run a longer experiment on a CLARIN-Emo dataset, finetuning a larger model, namely the large version of XLM-RoBERTa model from [8] using FocalLoss with weighted BCEWithLogitSLoss function as the cross-entropy part of FocalLoss, and included stochastic weight averaging.

Hyperparameter Settings. In the experiments using the Polish DistilRoBERTa language model on both datasets, the optimal learning rate was $1e-5$. We used AdamW optimizer with a weight decay of 0.01 and a learning rate scheduler with the fraction of the steps to increase the learning rate of 0.1. The training epoch numbers were obtained for each experiment scenario separately using early stopping with a patience of 20 epochs. The SSC multilabel task and the SSC UserID task on Clarin-Emo trained for 50 epochs, and for the SSC multilabel task on Stockbrief-GPT, it was 70 epochs. To prevent overfitting, a dropout layer with a rate of 0.1 was added. To train the larger model we use the same optimizer and its parameters alongside a linear schedule with a warmup scheduler with 10% warmup steps. We use early stopping of 10 epochs and do 20 annealing epochs for the stochastic weight averaging, with a learning rate of $1e-6$.

5.2 Text Classification

All of our datasets contain labels concerning text as a whole. We evaluate separately the model’s performance on all three of them. All three datasets have been split into the train (70%), dev (15%), and test (15%). In the case of CLARIN-Emo and Stockbrief, the documents in each split are the same as in sequential sentence classification.

Language model. To achieve at least somewhat comparable results all three models are fine-tuning the same model, the base version of XLM-RoBERTa. We chose a multilingual model for the text classification comparison as two of our datasets are in Polish and one is in English.

Hyperparameter Settings. Similarly to sequential sentence classification, we use AdamW optimizer with a learning rate of 1e-5 and weight decay of 0.01 and linear schedule with warmup scheduler with 10% warm up steps. For CLARIN-Emo and ChatGPT-Emo we find patience of 10 enough but we set it as 30 for StockBrief.

6 Results

Table 3 summarizes the results of the sequential sentence classification task on the CLARIN-Emo dataset. Figure 5 illustrates the results of the Polish DistilRoBERTa-based model on the Stockbrief-GPT dataset.

For the CLARIN-Emo dataset, regarding the macro F1-score, the personalized approach using UserID with Polish DistilRoBERTa surpassed the others, about 6 pp better than the XLM-RoBERTa-Large. However, concerning the micro F1-score, both models performed comparably, with 85.68% and 84.08% respectively. Overall, the UserID model achieved good performance on almost emotional labels, followed by the XLM-RoBERTa-Large, and in the final position is the SSC multilabel based on Polish DistilRoBERTa. *disgust* is the only emotion that Polish DistilRoBERTa got slightly higher than the other two models. Still, the difference is not significant, just about 2 pp. Besides, *sadness* is another emotion that UserID performed worse than the other two models, with a difference of about 6-7 pp. In addition, the lowest F1-scores were from two emotions, *surprise* and *fear* for all models. However, there is a notable difference between the highest (UserID) and the lowest (Polish DistilRoBERTa), around 24 pp and 11 pp for *surprise* and *fear*, respectively. Since these are two emotions with the lowest inter-annotator agreement and the lowest frequency in label distribution, it could explain this phenomenon and strengthen the benefit of personalized approaches.

Regarding the results from the Stockbrief-GPT dataset, while both micro F1-score (53.64%) and macro F1-score (47.72%) were not considered high and the lowest score being for *fear* (29.9%), it achieved considerably better scores on some emotions, such as *anticipation* and *joy*.

In the case of text classification, we achieved the highest results on ChatGPT-Emo (Figure 8). It achieves a macro-F1 score of 71.65%, with an abnormally high

Table 3: Sequential Sentence Multilabel Classification F1-score (%) on CLARIN-Emo.

F1-scores	PL-DistilRoBERTa	PL-DistilRoBERTa UserID	XLM-RoBERTa-Large
Micro	76.73	85.68	84.08
Macro	68.74	76.56	70.31
Joy	84.36	91.18	88.18
Trust	70.44	78.41	59.29
Anticipation	55.94	66.37	60.73
Surprise	34.14	58.85	45.19
Fear	41.97	52.82	43.04
Sadness	85.82	79.85	87.36
Disgust	67.17	64.57	64.96
Anger	65.22	69.63	68.06
Positive	86.20	92.06	88.80
Negative	87.15	93.00	90.04
Neutral	77.72	95.46	77.75

F1-score for *disappointed* label (98.05%). We achieve satisfactory results on the CLARIN-Emo dataset (Figure 6). As the text classification setup did not account for class balancing we can see the classifier struggle with rare labels but otherwise achieve a good performance. As presented in Figure 7, our classifier struggled on many of the emotion labels in Stockbrief-GPT, achieving only 22.86% macro F1-score. One of the possible explanations for this is too few training examples combined with a big class imbalance.

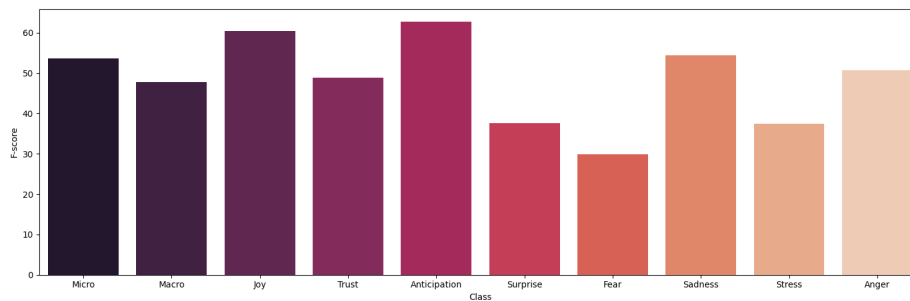


Fig. 5: Percentage F1-scores achieved in Sequential sentence multilabel classification on Stockbrief-GPT.

7 Conclusions and Future Work

In this paper, we analyzed several datasets annotated with emotions. The CLARIN-Emo dataset was the only dataset where both the text and the annotations

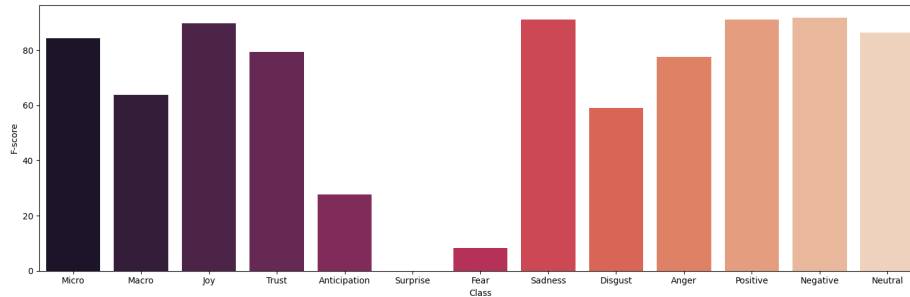


Fig. 6: Percentage F1-scores achieved in multilabel text classification on CLARIN-Emo.

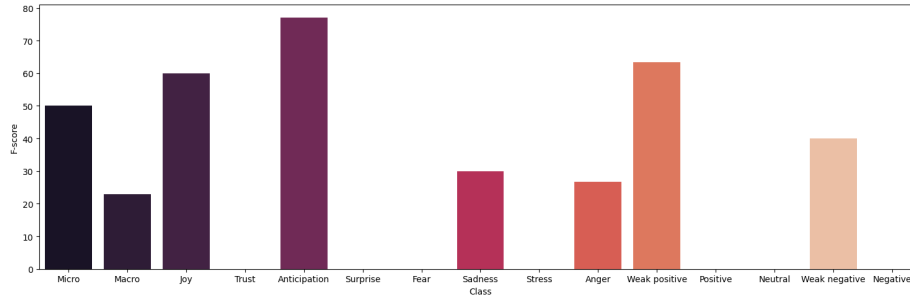


Fig. 7: Percentage F1-scores achieved in multilabel text classification on Stockbrief-GPT

came from a human. The ChatGPT-Emo dataset was generated entirely by the ChatGPT model, while the Stockbrief-GPT dataset contained text written by a human and annotated by ChatGPT. We acknowledge that the results are not directly comparable. At the same time, we note that emotion recognition is a difficult task. Manual annotation at the sentence level for the CLARIN-Emo set yielded a high-quality model trained on human labels (F1-macro at 68.74%). In paper [16], we showed that ChatGPT achieves only 53.23% on the same task (a decrease of about 15pp). In addition, due to the subjective nature of the task, we showed that adding human information to this set further increases F1-macro to 76.56% (an increase of about 8pp). If we train the model on the labels generated by the system at the sentence level, as in the case of the stock letter GPT set, we can obtain a model with an F1-macro quality of 47.72%, which requires further research as it remains to be seen what is the source of the uncertainty: the underdeterminism of the ChatGPT model or the lack of world knowledge to solve such a pragmatic task by other transformer-based models. Intuitively, however, one can assume that ChatGPT was not specifically trained to solve this type of task, hence the lower quality of both the annotation and the system trained on such an annotation.

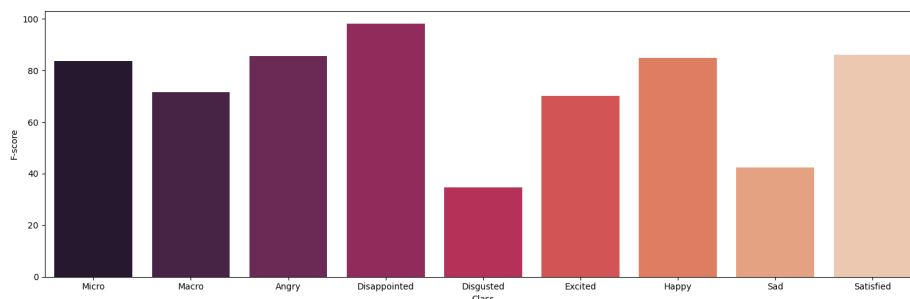


Fig. 8: Percentage F1-scores achieved in multilabel text classification on chatGPT-Emo.

The third dataset tested was built entirely with ChatGPT, i.e. both text and labels were generated by the model. Such a setup ensured a high consistency of emotion labeling for the text, yielding an F1-macro of 71.64% and an F1-micro of 83.71%. This is a much better result than the annotation evaluation performed by ChatGPT on human-written text. However, this may be due to the fact that the text generation prompt does not allow to enforce the diversity of the generated content, and many texts are semantically similar to each other. In further work, we plan to build a text generator based on random words from different domains, so that we can partially force a greater diversity of generated opinions.

References

1. Acheampong, F.A., Wenyu, C., Nunoo-Mensah, H.: Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports* **2**(7), e12189 (2020)
2. Antaki, F., Touma, S., Milad, D., El-Khoury, J., Duval, R.: Evaluating the performance of chatgpt in ophthalmology: An analysis of its successes and shortcomings. *medRxiv* (2023)
3. Baran, J., Kocoń, J.: Linguistic knowledge application to neuro-symbolic transformers in sentiment analysis. In: *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. pp. 395–402. IEEE (2022)
4. Bravo-Marquez, F., Frank, E., Mohammad, S.M., Pfahringer, B.: Determining word-emotion associations from tweets by multi-label classification. In: *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. pp. 536–539. IEEE (2016)
5. Chen, S.Y., Hsu, C.C., Kuo, C.C., Ku, L.W., et al.: Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379* (2018)
6. Chen, Y., Eger, S.: Transformers go for the lols: Generating (humorous) titles from scientific abstracts end-to-end (2022)
7. Cohan, A., Beltagy, I., King, D., Dalvi, B., Weld, D.: Pretrained language models for sequential sentence classification. 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. 3693–3699. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
8. 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. CoRR **abs/1911.02116** (2019)
 9. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: Goemotions: A dataset of fine-grained emotions. arXiv preprint arXiv:2005.00547 (2020)
 10. Desai, S., Caragea, C., Li, J.J.: Detecting perceived emotions in hurricane disasters. arXiv preprint arXiv:2004.14299 (2020)
 11. Ekman, P.: Are there basic emotions? *Psychological Review* **99**(3), 550–553 (1992)
 12. Gao, C.A., Howard, F.M., Markov, N.S., Dyer, E.C., Ramesh, S., Luo, Y., Pearson, A.T.: Comparing scientific abstracts generated by chatgpt to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv* (2022)
 13. Kazienko, P., Bielaniewicz, J., Gruza, M., Kanclerz, K., Karanowski, K., Miłkowski, P., Kocoń, J.: Human-centred neural reasoning for subjective content processing: Hate speech, emotions, and humor. *Information Fusion* (2023)
 14. Kim, Y., Jeong, S.R., Ghani, I.: Text opinion mining to analyze news for stock market prediction. *Int. J. Advance. Soft Comput. Appl* **6**(1), 2074–8523 (2014)
 15. Kocoń, J., Baran, J., Gruza, M., Janz, A., Kajstura, M., Kazienko, P., Korczyński, W., Miłkowski, P., Piasecki, M., Szolomicka, J.: Neuro-symbolic models for sentiment analysis. In: *Computational Science–ICCS 2022: 22nd International Conference, London, UK, June 21–23, 2022, Proceedings, Part II*. pp. 667–681. Springer (2022)
 16. Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieszczewski-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, L., Wojtasik, K., Woźniak, S., Kazienko, P.: Chatgpt: Jack of all trades, master of none. arXiv preprint arXiv:2302.10724 (2023)
 17. Kocoń, J., Gruza, M., Bielaniewicz, J., Grimling, D., Kanclerz, K., Miłkowski, P., Kazienko, P.: Learning personal human biases and representations for subjective tasks in natural language processing. In: *2021 IEEE International Conference on Data Mining (ICDM)*. pp. 1168–1173. IEEE (2021)
 18. Kocoń, J., Miłkowski, P., Kanclerz, K.: Multiemo: Multilingual, multilevel, multidomain sentiment analysis corpus of consumer reviews. In: *Computational Science–ICCS 2021: 21st International Conference, Krakow, Poland, June 16–18, 2021, Proceedings, Part II* 21. pp. 297–312. Springer (2021)
 19. 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)*. pp. 980–991 (2019)
 20. Korczyński, W., Kocoń, J.: Compression methods for transformers in multidomain sentiment analysis. In: *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. pp. 419–426. IEEE (2022)
 21. 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 (2019)
 22. Maynard, D.G., Greenwood, M.A.: Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In: *Lrec 2014 proceedings*. ELRA (2014)

23. Miłkowski, P., Gruza, M., Kazienko, P., Szolomicka, J., Woźniak, S., Kocoń, J.: Multi-model analysis of language-agnostic sentiment classification on multitemo data. In: Computational Collective Intelligence: 14th International Conference, ICCI 2022, Hammamet, Tunisia, September 28–30, 2022, Proceedings. pp. 163–175. Springer (2022)
24. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning based text classification: A comprehensive review (2020)
25. Mirzaee, H., Peymanfard, J., Moshtaghin, H.H., Zeinali, H.: Armanemo: A persian dataset for text-based emotion detection. arXiv preprint arXiv:2207.11808 (2022)
26. Ngo, A., Candri, A., Ferdinan, T., Kocoń, J., Korczynski, W.: Studemo: A non-aggregated review dataset for personalized emotion recognition. In: Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022. pp. 46–55 (2022)
27. Paul, M.J., Sarker, A., Brownstein, J.S., Nikfarjam, A., Scotch, M., Smith, K.L., Gonzalez, G.: Social media mining for public health monitoring and surveillance. In: Biocomputing 2016: Proceedings of the Pacific symposium. pp. 468–479. World Scientific (2016)
28. Păvăloaia, V.D., Teodor, E.M., Fotache, D., Danileț, M.: Opinion mining on social media data: sentiment analysis of user preferences. Sustainability **11**(16), 4459 (2019)
29. Phillips, T., Saleh, A., Glazewski, K.D., Hmelo-Silver, C.E., Mott, B., Lester, J.C.: Exploring the use of gpt-3 as a tool for evaluating text-based collaborative discourse. Companion Proceedings of the 12th p. 54 (2022)
30. Plutchik, R.: A general psychoevolutionary theory of emotion. In: Theories of emotion, pp. 3–33. Elsevier (1980)
31. Rashkin, H., Bosselut, A., Sap, M., Knight, K., Choi, Y.: Modeling naive psychology of characters in simple commonsense stories. arXiv preprint arXiv:1805.06533 (2018)
32. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (11 2020)
33. Rotsztein, J., Hollenstein, N., Zhang, C.: ETH-DS3Lab at SemEval-2018 task 7: Effectively combining recurrent and convolutional neural networks for relation classification and extraction. In: Proceedings of the 12th International Workshop on Semantic Evaluation. pp. 689–696. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018)
34. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (2019)
35. Sprugnoli, R., et al.: Multiemotions-it: A new dataset for opinion polarity and emotion analysis for italian. In: Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020). pp. 402–408. Accademia University Press (2020)
36. Susnjak, T.: Chatgpt: The end of online exam integrity? arXiv preprint arXiv:2212.09292 (2022)
37. Szolomicka, J., Kocon, J.: Multiaspectemo: Multilingual and language-agnostic aspect-based sentiment analysis. In: 2022 IEEE International Conference on Data Mining Workshops (ICDMW). pp. 443–450. IEEE (2022)
38. Tabone, W., de Winter, J.: Using chatgpt for human–computer interaction research: A primer. Manuscript submitted for publication (2023)