From Big to Small Without Losing It All: Text Augmentation with ChatGPT for Efficient Sentiment Analysis

1st Stanisław Woźniak
Department of Artificial Intelligence
Wrocław University of Science and Technology
Wrocław, Poland
stanislaw.wozniak@pwr.edu.pl

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

***Abstract*—In the era of artificial intelligence, data is gold but costly to annotate. The paper demonstrates a groundbreaking solution to this dilemma using ChatGPT for text augmentation in sentiment analysis. We leverage ChatGPT's generative capabilities to create synthetic training data that significantly improves the performance of smaller models, making them competitive with, or even outperforming, their larger counterparts. This innovation enables models to be both efficient and effective, thereby reducing computational cost, inference time, and memory usage without compromising on quality. Our work marks a key advancement in the cost-effective development and deployment of robust sentiment analysis models.**

***Index Terms*—Text Augmentation, ChatGPT, Sentiment Analysis, Model Efficiency, Data Annotation Cost**

I. Introduction

The burgeoning field of artificial intelligence (AI) is fueled by data—vast, often labeled datasets that serve as the training ground for machine learning models. Yet, collecting and annotating this data can be prohibitively expensive and time-consuming, creating a bottleneck in developing and deploying AI technologies. Parallel to these challenges, large language models like ChatGPT have attracted considerable attention for their remarkable generative capabilities, simulating human-like conversation and problem-solving skills [1]–[3].

In light of these developments, we explore a synergistic approach that leverages the generative prowess of ChatGPT to enhance the efficiency and performance of smaller machine-learning models in sentiment analysis tasks. Specifically, this paper investigates text augmentation as a data enhancement strategy in which synthetic samples are generated from existing data to supplement the training set.

The dual challenges of data sparsity and the high cost of human annotation have spurred interest in data augmentation techniques. However, previous approaches have often relied on simple heuristics or rule-based systems, limiting their applicability and effectiveness. We posit that large language models like ChatGPT can serve as robust, general-purpose engines for generating high-quality synthetic data, thereby alleviating these limitations.

We conduct comprehensive experiments with transformer-based models, such as RoBERTa, across two different sentiment analysis datasets: PerSenT [4] and MultiEmo [5]. The central questions we seek to address are twofold: First, can ChatGPT-augmented data boost the performance of smaller models to a level comparable with larger, more resource-intensive models? Second, do the benefits of using augmented data extend across different scales of models, from smaller to larger architectures?

By focusing on these questions, this paper aims to illuminate the potential of text augmentation via large language models, not only to improve model performance but also as a strategy for achieving computational efficiency and cost-effectiveness in AI applications.

Our list of contributions is as follows:

1) We present an innovative approach that exploits ChatGPT's generative abilities to enhance the training data for sentiment analysis tasks. This strategy is poised to circumvent the challenges associated with collecting and annotating vast datasets.
2) By leveraging ChatGPT for data augmentation, our methodology moves beyond the constraints of conventional heuristic and rule-based augmentation techniques, showcasing the potential for a more flexible and high-quality data enhancement.
3) We conduct experiments with transformer models, particularly RoBERTa, on two sentiment analysis datasets (PerSenT and MultiEmo). These experiments probe into the efficiency and performance of models trained on ChatGPT-augmented data.
4) The paper delves into pivotal questions concerning the scalability benefits of using augmented data and the feasibility of achieving high performance with smaller, less resource-intensive models.
5) Our research emphasizes improving models' performance and underscores the importance of achieving computational thriftiness and cost-effectiveness in AI applications, especially in a world leaning towards on-device deployments.

2375-9259/23/$31.00 ©2023 IEEE
DOI 10.1109/ICDMW60847.2023.00108

II. Related Work

A. Text Augmentation Techniques

The scarcity of data in machine learning tasks, particularly text classification, poses challenges for model training. A common remedy for this is text augmentation, where new data is artificially generated from the existing dataset to enrich the model's learning experience [6]–[8]. Techniques for text augmentation are diverse and can be applied at various levels of granularity, as outlined below:

- **Character Level:** Methods at this level manipulate individual characters to generate new text [9]–[11]. The advantage here is the subtlety of the changes, which can be nearly imperceptible yet effective in creating variations. However, the downside is that such minute alterations may sometimes lead to meaningless or garbled words, which may not aid in effective training.
- **Word Level:** This involves techniques like synonym replacement, word swapping, and more to augment text at the lexical level [12]–[20]. This method is beneficial as it maintains the overall sentence structure while introducing variability. The drawback, however, is that certain synonym replacements might change the context or nuance of the original sentence, which could be problematic for sensitive tasks.
- **Sentence Level:** Sentence-level augmentation modifies or restructures existing sentences for new variants [21]–[25]. Their strength lies in generating more diverse sentence structures, thereby potentially enriching the dataset. However, there's a risk that excessive modification might divert from the original intent or sentiment of the source sentence, making the augmentation counterproductive.
- **Document Level:** Entire documents are augmented to create data that maintains contextuality and coherence [26]–[28]. Their advantage is ensuring the augmented data still feels like a cohesive unit, which is particularly crucial for tasks that rely on understanding context over longer passages. On the flip side, these methods might be computationally more intensive and challenging to implement as maintaining contextuality across a full document can be complex.

In the present study, our primary focus is on document-level text augmentation.

B. Advancements in Large Language Models (LLMs)

Transformers have revolutionized the field of natural language processing [29]. Specifically, generative transformer models, known as large language models (LLMs), are becoming increasingly popular owing to their extensive parameter space [30]. Several variants exist:

- **GPT Series:** Including GPT-3 and GPT-4, these models are benchmarks in the field [31].
- **FLAN and Bloom:** These models represent advancements in specific aspects like efficiency and capability [32], [33].
- **LLaMA Family:** Known for their robustness across multiple tasks [34], [35].
- **ChatGPT:** A specialized version of InstructGPT trained via reinforcement learning from human feedback (RLHF), known for conversational applications [3], [36].

This work will employ ChatGPT due to its adaptability and general-purpose capabilities.

C. Data Generation and Augmentation via LLMs

LLMs offer an intriguing avenue for generating and augmenting data. These models can produce data that closely aligns with human-generated content by feeding specific prompts. Some studies have utilized ChatGPT to generate entire synthetic datasets [37], [38]. Others have explored using LLMs for augmenting existing datasets [39]–[42]. These methods have generally led to improved performance in various machine-learning tasks.

D. Deep Learning in Sentiment Analysis

Sentiment analysis, the focus of this paper, is a well-studied problem aimed at classifying the sentiment conveyed in a text. Over the years, many methods and architectures have been proposed and implemented to decipher and categorize sentiments in textual data [5], [43]–[56]. Ranging from early lexicon-based techniques to advanced pre-trained models, the field has seen vast developments.

- **Lexicon-based Techniques:** These techniques rely on sentiment dictionaries or lexicons, wherein words are pre-associated with sentiment scores. Tools such as SenticNet have further enhanced the lexicon-based approaches, combining commonsense reasoning with neuro-symbolic computations for a more holistic sentiment analysis [57].
- **Recurrent Neural Networks:** Models such as BiLSTM/CNN and MCBiGRU are adept at capturing text-dependent sequential dependencies. Their ability to memorize previous information and context makes them suitable for sentiment analysis tasks [58]. A newer trend in sentiment analysis and comment toxicity detection, this method employs a combination of convolutional and recurrent architectures to process text from multiple channels, ensuring a more comprehensive understanding of the data [59].
- **Transformer Models:** Techniques like BERT and RoBERTa have risen in prominence due to their unparalleled capability to capture contextual information. These architectures, especially when pre-trained, have set new standards in sentiment analysis benchmarks [44], [60]–[64]. An increasingly popular choice, these models, trained on vast amounts of text data, possess an inherent understanding of language semantics and structure. However, concerns regarding biases in these models have been highlighted in recent literature, emphasizing the need for careful and context-aware implementation [49], [65]–[75].

Incorporating these diverse methods, this paper aims to make significant contributions to the field of sentiment analysis. A particular emphasis is placed on leveraging ChatGPT for data augmentation, bridging the gap between deep learning techniques and real-world sentiment analysis applications.

III. Data Augmentation Methodology

To augment our dataset, we utilized OpenAI's GPT-3.5 model via its API, employing default settings. Our method involves generating new data points based on the original dataset by using four distinct prompts. These prompts are designed to cover two different approaches to text augmentation: paraphrasing and generating entirely new text inspired by the original.

A. Prompt-based Strategy

Our approach adopts a two-tier structure for each augmentation method: *paraphrasing* and *inspirational generation*. However, it's essential to understand the differences and potential implications of each method:

- **Paraphrasing:** The primary goal of paraphrasing is to offer a varied representation of the same message or content. While it brings subtle variations to the text, it often doesn't exceed the inherent domain or context of the original content. This limitation can be both an advantage and a drawback. The advantage is that the paraphrased content remains tightly related to the original data, ensuring that the augmented data remains contextually relevant. However, the downside is that this might not sufficiently diversify the dataset, potentially limiting the enhanced generalization of the model.
- **Inspirational Generation:** This strategy aims to craft content that is distinct from the original but maintains the same sentiment. The idea is to widen the scope of data to potentially traverse different domains or contexts, which might provide a more comprehensive augmentation. This method's primary advantage is that it might aid in creating a more domain-agnostic model, as the new texts could span diverse themes. However, there's a risk involved. As the generated content deviates from the original, there's a possibility that some generated examples might not align perfectly with the expected sentiment, potentially introducing noise into the dataset.

B. Detailed Prompt Descriptions

For both augmentation strategies, we've delineated the process with specific prompts:

- For **Paraphrasing:** The first two prompts come under this umbrella. They are presented sequentially in a single conversation session with the model:
  - "Generate a paraphrase for the following text, preserving the sentiment of the following statement: *text*"
  - "Generate another paraphrase by changing more words also keeping the sentiment"
- For **Inspirational Generation:** The latter two prompts are crafted to generate sentiment-consistent yet distinct content. They are initiated in individual conversation sessions with the model:
  - "Based on the given text, generate another text with a completely new theme, but be inspired by the original text and keep the sentiment of the old one in the new text. Original text: *text*"
  - "Based on the given text, generate another text with a completely new theme, but be inspired by the original text and keep the *label* sentiment. Original text: *text*"

C. Implementation Details

For every sample in each dataset, we applied all four prompts. It is important to note that:

- A single session was used for the first two prompts, wherein the second prompt was input only after the model responded to the first.
- The third and fourth prompts were processed in separate sessions. The primary distinction between them is that the third prompt directs the model to infer the sentiment from the original text, whereas the fourth one includes the sentiment label explicitly in the prompt.

D. Output Datasets

We generated four separate augmented datasets by applying these prompts to each sample from the original training set. Each of these datasets maintains the class distribution of the original data, providing us with a richer resource for training our machine learning models.

Example augmentation
Sample from the original dataset
With my children (10 years old and 2 years old) I actually only go to doctor. I think that she is a great specialist, she does not treat the patient "from above", she is nice, factual and well remembers and knows her little patients. She has an excellent attitude towards children and they like her very much. She does not prescribe immediately, sometimes without any sense, an antibiotic, only when it is necessary. A few times I happened to get to another pediatrician, but as a rule I was not satisfied. I'm glad that the doctor is working in my clinic. When he leaves, I'll have trouble choosing another pediatrician. I highly recommend it!
Paraphrased sample by the first prompt
I only take my children (10 years old and 2 years old) to one doctor, and I believe she is an exceptional specialist. She doesn't treat her patients with condescension and has a warm and friendly demeanor. She demonstrates excellent knowledge of and familiarity with her young patients. It is evident that she genuinely cares about children and they have a strong liking for her. She only prescribes antibiotics when necessary and avoids unnecessary medication. Although I have occasionally visited other pediatricians, I have mostly been dissatisfied. I am grateful that this doctor is a part of my clinic, and I anticipate difficulty in choosing a new pediatrician when she eventually leaves. I highly recommend her!
Generated sample by the third prompt

When it comes to planning family vacations, I usually rely on only one travel agency. Not just any agency, but one that I consider to be exceptional in providing personalized service and creating unforgettable experiences. They don't treat their clients as mere customers, but as cherished guests, ensuring every aspect of our trip is taken care of. From the moment we step foot into their office, their friendly and knowledgeable staff greet us with warmth and empathy. They have a remarkable understanding of our preferences and take note of even the smallest details that truly matter to us. Unlike other agencies, they don't rush into booking the first available package. Instead, they carefully consider our desires and requirements before suggesting the perfect destination and itinerary. They genuinely care about our well-being and never recommend unnecessary add-ons or impulse purchases. I've occasionally tried other agencies, but I have always been left dissatisfied. It's the kind of satisfaction that lingers in your mind, making it difficult to choose another agency in the future. I am truly grateful for the exceptional team at this agency. Their dedication, professionalism, and genuine love for what they do sets them apart. It's not just about business for them; it's about creating magical memories for families like mine. I can wholeheartedly recommend their services to anyone searching for a travel agency that goes above and beyond.
Sentiment label
Positive

IV. Datasets

For the sentiment analysis classification task, this paper employs two distinct datasets: PerSenT [4] and MultiEmo [5]. Both datasets are well-suited for our research aims but differ in document length, language diversity, and class labels. A detailed statistical summary of each dataset is provided to understand their characteristics comprehensively.

A. PerSenT

The PerSenT dataset encompasses approximately 50,000 documents. On average, each document in this dataset comprises 377 words. The dataset is divided into three subsets: a training set with 3,355 documents, a validation set containing 578 documents, and a test set with 827 documents.

Texts in PerSenT are labeled according to one of three sentiment classes: *Positive*, *Negative*, and *Neutral*. The distribution of these classes across the training, validation, and test sets can be found in Table I.

B. MultiEmo

MultiEmo is a more diverse dataset constructed from consumer reviews across four different domains: medicine, hotels, products, and universities. The entire corpus consists of documents in 11 languages, and each language subset contains roughly 8,000 documents. For the scope of this paper, we focus solely on the English documents combined from all domains.

In terms of document length, MultiEmo texts are generally shorter than those in PerSenT, averaging around 140 words. The dataset is divided into training, validation, and test subsets, containing 6,572, 823, and 820 texts respectively. Unlike PerSenT, MultiEmo includes a fourth sentiment class, *Ambivalent*, along with *Positive*, *Negative*, and *Neutral*. The class distribution is detailed in Table I.

V. Experiments

A. Setup

The experiments were conducted using three transformer-based models: XLM-RoBERTa-comet-small [76] (referred to as RoBERTa-small), XLM-RoBERTa-base [77] (referred to as RoBERTa-base), and XtremeDistil [78], [79] configured with six layers and a hidden size of 384. The parameter counts for these models are as follows:

- RoBERTa-base: 279 million parameters
- RoBERTa-small: 107 million parameters
- XtremeDistil: 13 million parameters

1) Data Augmentation Strategies: We employed various original and augmented data combinations to form new training sets. Seven unique combinations were tested, each preserving the original class distribution. For every augmentation strategy and specific prompt (Section III-B), we generated synthetic data equivalent to the volume of the original training set.

2) Baseline Experiment: The baseline experiment used only the original training set without any form of data augmentation. All other experiments extended the original training set with augmented data generated using specific prompts, as detailed below:

- *Para*: Contains original training set and data augmented using the first prompt for paraphrasing.
- *Para-Conv*: Contains original training set and data augmented using the second prompt for paraphrasing in a conversational context.
- *Both Para*: Merges *Para* and *Para-Conv*.
- *Insp*: Contains original training set and data augmented using the third prompt for generating inspired text.
- *Insp-Lab*: Similar to *Insp*, but uses the fourth prompt that includes sentiment labels.
- *Both Insp*: Merges *Insp* and *Insp-Lab*.
- *All*: Consolidates all augmented data with the original training set.

3) Model Training Parameters: All models were trained with a learning rate of 10^{-5} – a typical value often adopted when fine-tuning models, especially for transformer-based architectures [61]. The batch size was set to 16, a decision primarily influenced by the memory limitations of our graphics card (RTX 3090). Tokenizers were specifically configured to truncate texts, capping them at a maximum of 512 tokens. This truncation length was selected because 512 tokens represent our model's maximum allowable input length.

B. Experimental Results

With the seven data combinations and three transformer models, 42 different experiments were conducted. Each experiment used the validation and test sets from the original data splits without modification. Each experimental run was performed five times with varying random seeds for reliability. The code is available on the GitHub repository[1].

[1]https://github.com/CLARIN-PL/text-augumentation-with-chatgpt

TABLE I
THE DATASETS CLASS DISTRIBUTION

Dataset split	PerSenT			MultiEmo			
	Positive	*Negative*	*Neutral*	*Positive*	*Negative*	*Neutral*	*Ambivalent*
train	52.4%	10.46%	37.14%	27.74%	37.57%	14.77%	19.92%
valid	52.6%	10.03%	37.37%	28.68%	36.94%	15.55%	18.83%
test	44.5%	16.81%	38.69%	27.68%	41.34%	14.39%	16.59%

C. Inference Time Evaluation

To assess the inference efficiency of each model, we conducted a separate experiment measuring the time required for processing a sample batch of size 16 across 2,000 iterations. Average inference times for each model were calculated for both the PerSenT and MultiEmo datasets, as detailed in section IV.

D. Results

1) Evaluation Metrics: Our evaluation of the experiments is based on three primary metrics: accuracy, F1 macro, and gain [3]. The gain metric quantifies the improvement of a given model relative to the baseline model for a specific metric, either accuracy or F1 score. It is calculated as shown in equation 1.

$$\text{Gain} = \frac{100\% \times (M - B)}{100\% - B} \tag{1}$$

2) PerSenT Dataset Results: Results obtained from the PerSenT dataset are summarized in Table II. These results show the F1 macro metric and accuracy averaged over 5 repetitions, alongside standard deviations. Generally, the augmented datasets yielded superior results compared to the baselines.

TABLE II
RESULTS PERSENT

Augmentation Type	Transformer		
	RoBERTa-small	**RoBERTa-base**	**XtremDistil**
	F1 macro		
Baseline	$36\% \pm 2\%$	$38\% \pm 8\%$	$41\% \pm 3\%$
Para	$38\% \pm 3\%$	$40\% \pm 1\%$	$41\% \pm 2\%$
Para-Conv	$39\% \pm 1\%$	$41\% \pm 1\%$	$43\% \pm 2\%$
Both Para	$40\% \pm 2\%$	$37\% \pm 1\%$	$43\% \pm 2\%$
Insp	$37\% \pm 4\%$	$43\% \pm 1\%$	$43\% \pm 3\%$
Insp-Lab	$38\% \pm 2\%$	$41\% \pm 2\%$	$40\% \pm 2\%$
Both Insp	$37\% \pm 1\%$	$41\% \pm 2\%$	$42\% \pm 3\%$
All	$39\% \pm 2\%$	$39\% \pm 1\%$	$43\% \pm 1\%$
	Accuracy		
Baseline	$38\% \pm 2\%$	$39\% \pm 8\%$	$43\% \pm 5\%$
Para	$39\% \pm 4\%$	$46\% \pm 1\%$	$43\% \pm 3\%$
Para-Conv	$39\% \pm 1\%$	$44\% \pm 1\%$	$44\% \pm 2\%$
Both Para	$41\% \pm 3\%$	$45\% \pm 3\%$	$45\% \pm 3\%$
Insp	$37\% \pm 5\%$	$46\% \pm 1\%$	$44\% \pm 3\%$
Insp-Lab	$41\% \pm 2\%$	$46\% \pm 4\%$	$42\% \pm 2\%$
Both Insp	$40\% \pm 1\%$	$46\% \pm 2\%$	$44\% \pm 3\%$
All	$44\% \pm 2\%$	$46\% \pm 2\%$	$46\% \pm 1\%$

3) MultiEmo Dataset Results: Metrics from the MultiEmo dataset are presented in Table III. Though the dataset is less complex than PerSenT, we still observed improvements when using augmented data.

TABLE III
RESULTS MULTIEMO

Augmentation Type	Transformer		
	RoBERTa-small	**RoBERTa-base**	**XtremDistil**
	F1 macro		
Baseline	$78\% \pm 2\%$	$87\% \pm 1\%$	$85\% \pm 1\%$
Para	$83\% \pm 0\%$	$88\% \pm 1\%$	$86\% \pm 1\%$
Para-Conv	$82\% \pm 1\%$	$87\% \pm 1\%$	$87\% \pm 1\%$
Both Para	$83\% \pm 1\%$	$87\% \pm 1\%$	$86\% \pm 0\%$
Insp	$82\% \pm 1\%$	$86\% \pm 1\%$	$85\% \pm 0\%$
Insp-Lab	$82\% \pm 0\%$	$87\% \pm 0\%$	$85\% \pm 1\%$
Both Insp	$84\% \pm 1\%$	$86\% \pm 1\%$	$85\% \pm 0\%$
All	$85\% \pm 1\%$	$86\% \pm 1\%$	$87\% \pm 1\%$
	Accuracy		
Baseline	$80\% \pm 1\%$	$88\% \pm 1\%$	$86\% \pm 1\%$
Para	$84\% \pm 1\%$	$89\% \pm 1\%$	$87\% \pm 1\%$
Para-Conv	$83\% \pm 1\%$	$88\% \pm 1\%$	$87\% \pm 1\%$
Both Para	$84\% \pm 1\%$	$88\% \pm 1\%$	$80\% \pm 0\%$
Insp	$84\% \pm 1\%$	$87\% \pm 1\%$	$86\% \pm 0\%$
Insp-Lab	$84\% \pm 1\%$	$88\% \pm 1\%$	$86\% \pm 1\%$
Both Insp	$84\% \pm 1\%$	$87\% \pm 1\%$	$86\% \pm 0\%$
All	$85\% \pm 1\%$	$88\% \pm 1\%$	$88\% \pm 1\%$

4) Comparative Gain Analysis: The gain metric, calculated per equation 1, was computed for all experiments. Figure 1 and Figure 2 illustrate the gains for PerSenT and MultiEmo datasets, respectively.

In the less complex MultiEmo dataset, larger models often exhibited negative gains, while smaller models displayed positive gains. However, in the more complex PerSenT dataset, RoBERTa-base showed substantial improvements over the baseline.

5) Class-Level Gain Analysis: We further delved into the gain metric at the class level. Figures 3, 4, 5, and 6 display these metrics for both datasets. Interesting observations were made, such as significant improvements in the *Ambivalent* class with RoBERTa-small on the MultiEmo dataset, with little to no trade-offs in other classes.

6) Comparison with Baseline: Figures 7 and 8 contrast the best-performing models trained on augmented datasets against their corresponding baselines. In almost all cases, augmented training data improved model performance, bringing smaller models close to or surpassing the performance of larger models trained on original data.

7) Inference Time Analysis: Inference time results are presented in Figure 9. As observed, RoBERTa-base had the longest inference times—almost seven times slower than XtremeDistil. It is worth noting that the time metrics were consistent across datasets, even though the MultiEmo texts were considerably shorter on average, thus confirming text

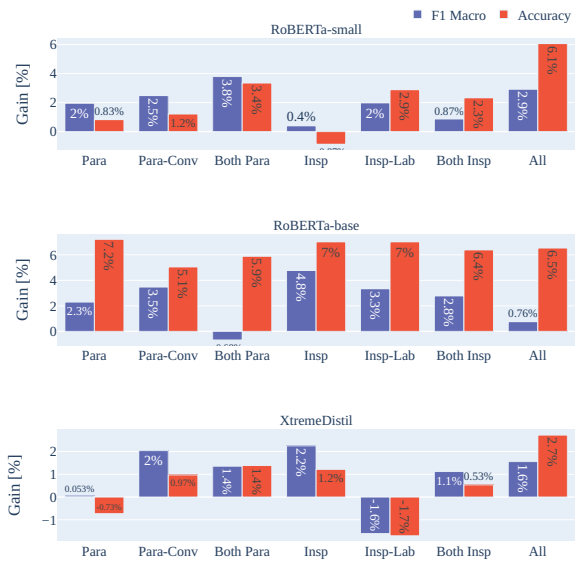

Fig. 1. Gain metric per augmentation type on PerSenT dataset in accuracy and F1 macro metrics

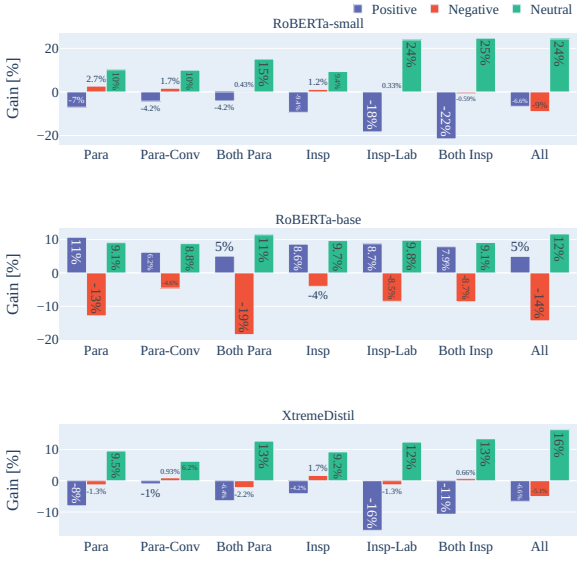

Fig. 3. Gain metric per class and augmentation type on PerSenT dataset in F1 macro

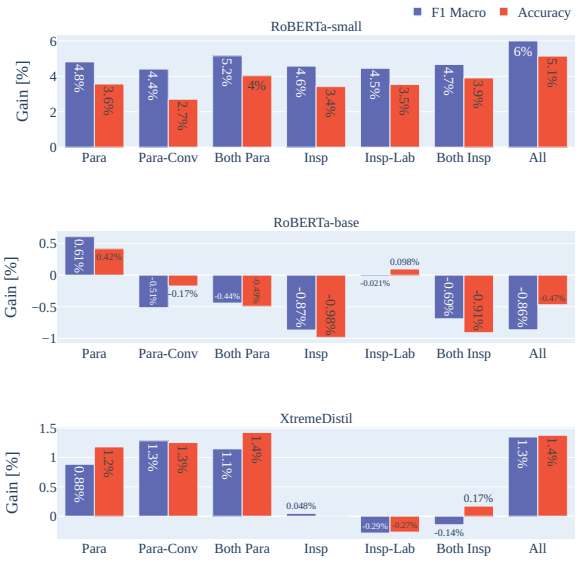

Fig. 2. Gain metric per augmentation type on MultiEmo dataset in accuracy and F1 macro metrics

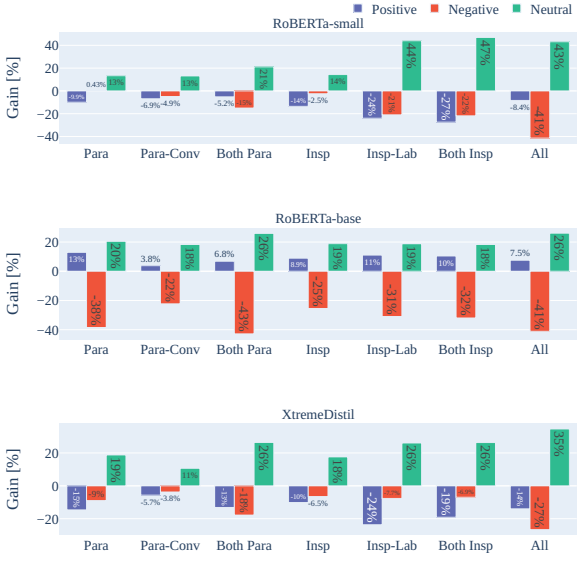

Fig. 4. Gain metric per class and augmentation type on PerSenT dataset in accuracy

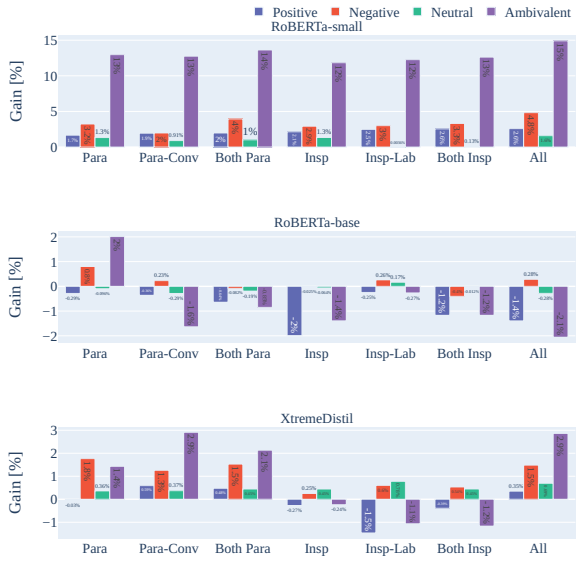

Fig. 5. Gain metric per class and augmentation type on MultiEmo dataset in F1 macro

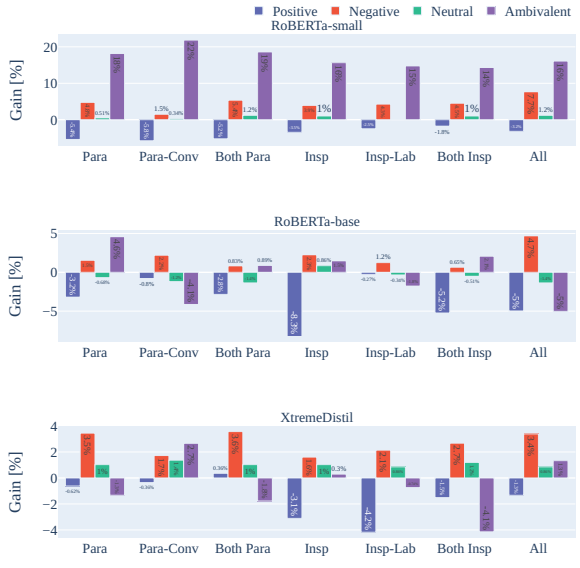

Fig. 6. Gain metric per class and augmentation type on MultiEmo dataset in accuracy

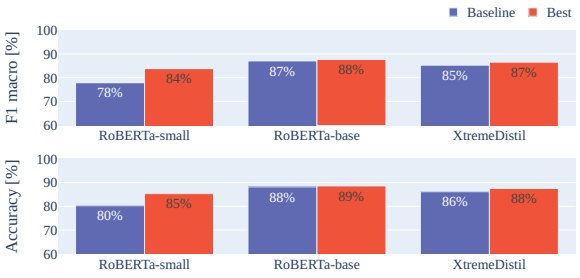

Fig. 7. Baseline vs. best model on Multiemo dataset

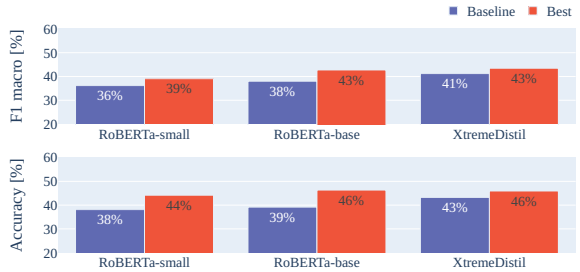

Fig. 8. Baseline vs. best model on PerSenT dataset

length as a variable influencing inference time.

E. Conclusions

1) Data Augmentation Efficacy: The results affirm that data augmentation, specifically using large language models (LLMs), enhances the performance of transformer models in the tested tasks [76]–[78]. However, the effectiveness of augmentation varied depending on the type and the dataset in question.

2) Optimization of Resources: Our experiments highlight an important trade-off between computational resources and model performance. Training larger models, particularly with

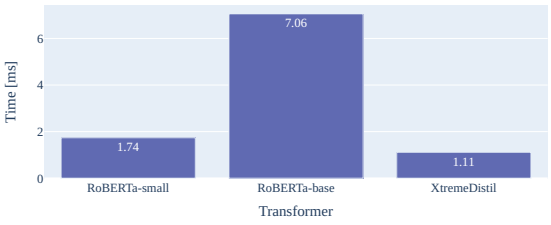

Fig. 9. Inference time per sample

augmented data, can be resource-intensive and costly. Yet, our results suggest that smaller models trained on augmented datasets can achieve competitive or even superior performance while substantially reducing computational requirements [76], [78].

3) Performance Efficiency: Smaller models benefited from speed-ups of nearly 7x during inference and saved up to 20 times the memory in terms of the number of parameters without compromising on performance quality. This suggests that adopting smaller, more efficient models is a viable strategy, especially when coupled with augmented training data [76], [78].

4) Class-Level Performance: An intriguing observation was the differential impact of data augmentation on class-level performance. Larger models, particularly when trained on augmented datasets, demonstrated more significant performance degradation in specific classes without commensurate gains in other classes. This finding strengthens the argument for the utility of smaller models trained on augmented data, as they tend to offer a more balanced improvement across classes [3].

5) Future Directions: While our work lays a foundational understanding of the benefits of data augmentation and model selection, ample room remains for exploration. Future work could investigate the effects of more nuanced forms of augmentation or delve deeper into the relationship between augmentation and the class distribution in datasets.

6) Overall Summary: In summary, our research demonstrates that strategic data augmentation can improve model performance while mitigating the challenges associated with resource-intensive larger models. This study paves the way for more efficient and cost-effective machine learning models, highlighting the merits of smaller, augmented models over their larger counterparts.

VI. Discussion and Future Work

As the realm of artificial intelligence continues to expand, the imperative for deploying models on diverse platforms, including resource-constrained devices such as mobile phones, becomes increasingly salient. This paper has demonstrated that it is possible to maintain high-performance levels with significantly smaller models, thereby enabling their use on devices with limited memory without compromising the quality of results. Furthermore, the compact nature of these models enables on-device inference, thereby augmenting user privacy and security, which are critical concerns in today's digital landscape.

In the present study, we have concentrated on data augmentation utilizing GPT-3.5 as the large language model (LLM) of choice. Moving forward, several avenues for exploration and enhancement present themselves. One such avenue would be to investigate other LLMs for their potential in augmenting data. Alternative LLMs could offer superior capabilities in sentiment prediction, thereby enhancing the quality of the augmented dataset. Furthermore, some models might offer free API access without sacrificing quality, which could reduce costs substantially. The speed of the chosen model could also be a significant factor, as faster models would allow for quicker data collection, leading to potential cost savings.

Another area ripe for future inquiry is the optimization of prompt engineering for the selected LLM. Crafting more efficient and targeted prompts could yield augmented data of even higher quality. Beyond mere paraphrasing or simple data extensions, the use of carefully designed prompts could produce a broader range of text augmentations. These might include more nuanced reinterpretations of the original data, thus expanding the scope and diversity of the generated dataset. Additionally, more efficient prompt engineering could produce multiple, high-quality augmentations within a single model response, thereby enriching the dataset without a corresponding increase in processing time.

Lastly, we propose future experiments that explore the integration of original training sets with their augmented counterparts, specifically for class balancing. By merging the two types of datasets, it might be possible to eliminate the issues of class imbalance, thereby yielding more reliable and generalizable results at the class level.

In summary, the work presented herein opens up several promising avenues for future research, ranging from LLM selection and prompt engineering to dataset integration for class balancing. These endeavors could not only improve the efficiency and cost-effectiveness of utilizing LLMs for data augmentation but also significantly advance the field's understanding of optimizing model performance on resource-constrained platforms.

Acknowledgements

This work was financed by (1) contribution to the European Research Infrastructure "CLARIN ERIC - European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure", 2022-23 (CLARIN Q); (2) the European Regional Development Fund, as a part of the 2014-2020 Smart Growth Operational Programme, projects no. POIR.04.02.00-00C002/19, POIR.01.01.01-00-0923/20, POIR.01.01.01-00-0615/21, and POIR.01.01.01-00-0288/22; (3) the statutory funds of the Department of Artificial Intelligence, Wroclaw University of Science and Technology; (4) the Polish Ministry of Education and Science within the programme "International Projects Co-Funded"; (5) the European Union under the Horizon Europe, grant no. 101086321 (OMINO). However, the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor European Research Executive Agency can be held responsible for them.

References

[1] R. Mao, G. Chen, X. Zhang, F. Guerin, and E. Cambria, "Gpteval: A survey on assessments of chatgpt and gpt-4," *arXiv preprint arXiv:2308.12488*, 2023.

[2] M. M. Amin, E. Cambria, and B. W. Schuller, "Will affective computing emerge from foundation models and general ai? a first evaluation on chatgpt," *IEEE Intelligent Systems*, vol. 38, p. 2, 2023.

[3] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz *et al.*, "Chatgpt: Jack of all trades, master of none," *Information Fusion*, p. 101861, 2023.

[4] M. Bastan, M. Koupaee, Y. Son, R. Sicoli, and N. Balasubramanian, "Author's sentiment prediction," *arXiv preprint arXiv:2011.06128*, 2020.

[5] J. Kocoń, P. Miłkowski, and K. Kanclerz, "Multiemo: Multilingual, multilevel, multidomain sentiment analysis corpus of consumer reviews," in *International Conference on Computational Science*. Springer, 2021, pp. 297–312.

[6] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for nlp," *arXiv preprint arXiv:2105.03075*, 2021.

[7] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning," *Journal of big Data*, vol. 8, pp. 1–34, 2021.

[8] M. Bayer, M.-A. Kaufhold, and C. Reuter, "A survey on data augmentation for text classification," *ACM Computing Surveys*, vol. 55, no. 7, pp. 1–39, 2022.

[9] Y. Belinkov and Y. Bisk, "Synthetic and natural noise both break neural machine translation," *arXiv preprint arXiv:1711.02173*, 2017.

[10] C. Coulombe, "Text data augmentation made simple by leveraging nlp cloud apis," *arXiv preprint arXiv:1812.04718*, 2018.

[11] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, "Hotflip: White-box adversarial examples for text classification," *arXiv preprint arXiv:1712.06751*, 2017.

[12] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," *arXiv preprint arXiv:1901.11196*, 2019.

[13] O. Kolomiyets, S. Bethard, and M.-F. Moens, "Model-portability experiments for textual temporal analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, vol. 2. ACL; East Stroudsburg, PA, 2011, pp. 271–276.

[14] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database," *International journal of lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[15] J. Kocoń, A. Janz, and M. Piasecki, "Context-sensitive sentiment propagation in wordnet," in *Proceedings of the 9th global wordnet conference*, 2018, pp. 329–334.

[16] ——, "Classifier-based polarity propagation in a wordnet," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[17] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," *arXiv preprint arXiv:1805.06201*, 2018.

[18] J. Kocoń and A. Janz, "Propagation of emotions, arousal and polarity in wordnet using heterogeneous structured synset embeddings," in *Proceedings of the 10th Global Wordnet Conference*, 2019, pp. 336–341.

[19] M. Wierzba, M. Riegel, J. Kocoń, P. Miłkowski, A. Janz, K. Klessa, K. Juszczyk, B. Konat, D. Grimling, M. Piasecki *et al.*, "Emotion norms for 6000 polish word meanings with a direct mapping to the polish wordnet," *Behavior Research Methods*, pp. 1–16, 2021.

[20] R. Mao, X. Li, K. He, M. Ge, and E. Cambria, "Metapro online: A computational metaphor processing online system," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 2023, pp. 127–135.

[21] S. Y. Feng, A. W. Li, and J. Hoey, "Keep calm and switch on! preserving sentiment and fluency in semantic text exchange," *arXiv preprint arXiv:1909.00088*, 2019.

[22] J. Min, R. T. McCoy, D. Das, E. Pitler, and T. Linzen, "Syntactic data augmentation increases robustness to inference heuristics," *arXiv preprint arXiv:2004.11999*, 2020.

[23] H. H. Kim, D. Woo, S. J. Oh, J.-W. Cha, and Y.-S. Han, "Alp: Data augmentation using lexicalized pcfgs for few-shot text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10 894–10 902.

[24] V. Marivate and T. Sefara, "Improving short text classification through global augmentation methods," in *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*. Springer, 2020, pp. 385–399.

[25] M. Pogoda, K. Gawron, N. Ropiak, M. Swędrowski, and J. Kocoń, "Deep neural sequence to sequence lexical substitution for the polish language," in *International Conference on Computational Science*. Springer, 2022, pp. 692–705.

[26] M. Aiken and M. Park, "The efficacy of round-trip translation for mt evaluation," *Translation Journal*, vol. 14, no. 1, pp. 1–10, 2010.

[27] S. Qiu, B. Xu, J. Zhang, Y. Wang, X. Shen, G. De Melo, C. Long, and X. Li, "Easyaug: An automatic textual data augmentation platform for classification tasks," in *Companion Proceedings of the Web Conference 2020*, 2020, pp. 249–252.

[28] X. Sun and J. He, "A novel approach to generate a large scale of supervised data for short text sentiment analysis," *Multimedia Tools and Applications*, vol. 79, no. 9-10, pp. 5439–5459, 2020.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[30] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[31] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[32] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei *et al.*, "The flan collection: Designing data and methods for effective instruction tuning," *arXiv preprint arXiv:2301.13688*, 2023.

[33] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé *et al.*, "Bloom: A 176b-parameter open-access multilingual language model," *arXiv preprint arXiv:2211.05100*, 2022.

[34] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models. corr, abs/2302.13971, 2023. doi: 10.48550," *arXiv preprint arXiv.2302.13971*.

[35] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[36] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[37] S. Ubani, S. O. Polat, and R. Nielsen, "Zeroshotdataaug: Generating and augmenting training data with chatgpt," *arXiv preprint arXiv:2304.14334*, 2023.

[38] W. Yang and G. Nicolai, "Neural machine translation data generation and augmentation using chatgpt," *arXiv preprint arXiv:2307.05779*, 2023.

[39] S. Sarker, L. Qian, and X. Dong, "Medical data augmentation via chatgpt: A case study on medication identification and medication event classification," *arXiv preprint arXiv:2306.07297*, 2023.

[40] H. Dai, Z. Liu, W. Liao, X. Huang, Z. Wu, L. Zhao, W. Liu, N. Liu, S. Li, D. Zhu *et al.*, "Chataug: Leveraging chatgpt for text data augmentation," *arXiv preprint arXiv:2302.13007*, 2023.

[41] C. Whitehouse, M. Choudhury, and A. F. Aji, "Llm-powered data augmentation for enhanced crosslingual performance," *arXiv preprint arXiv:2305.14288*, 2023.

[42] R. Y. Maragheh, L. Morishetti, R. Giahi, K. Nag, J. Xu, J. Cho, E. Korpeoglu, S. Kumar, and K. Achan, "Llm-based aspect augmentations for recommendation systems," 2023.

[43] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "Affective computing and sentiment analysis," *A practical guide to sentiment analysis*, pp. 1–10, 2017.

[44] J. Kocoń, M. Zaśko-Zielińska, and P. Miłkowski, "Multi-level analysis and recognition of the text sentiment on the example of consumer opinions," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 2019, pp. 559–567.

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

[46] J. Kocoń, P. Miłkowski, M. Wierzba, B. Konat, K. Klessa, A. Janz, M. Riegel, K. Juszczyk, D. Grimling, A. Marchewka *et al.*, "Multilingual and language-agnostic recognition of emotions, valence and arousal in large-scale multi-domain text reviews," in *Language and Technology Conference*. Springer, 2019, pp. 214–231.

[47] J. Kocon, A. Janz, P. Miłkowski, M. Riegel, M. Wierzba, A. Marchewka, A. Czoska, D. Grimling, B. Konat, K. Juszczyk *et al.*, "Recognition of

emotions, valence and arousal in large-scale multi-domain text reviews," *Human language technologies as a challenge for computer science and linguistics*, pp. 274–280, 2019.

[48] K. Kanclerz, P. Miłkowski, and J. Kocoń, "Cross-lingual deep neural transfer learning in sentiment analysis," *Procedia Computer Science*, vol. 176, pp. 128–137, 2020.

[49] J. Kocoń, A. Figas, M. Gruza, D. Puchalska, T. Kajdanowicz, and P. Kazienko, "Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach," *Information Processing & Management*, vol. 58, no. 5, p. 102643, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457321001333

[50] J. Kocoń and M. Maziarz, "Mapping wordnet onto human brain connectome in emotion processing and semantic similarity recognition," *Information Processing & Management*, vol. 58, no. 3, p. 102530, 2021.

[51] J. Baran and J. Kocoń, "Linguistic knowledge application to neuro-symbolic transformers in sentiment analysis," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2022, pp. 395–402.

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

[53] J. Szołomicka and J. Kocon, "Multiaspectemo: Multilingual and language-agnostic aspect-based sentiment analysis," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2022, pp. 443–450.

[54] W. Korczyński and J. Kocoń, "Compression methods for transformers in multidomain sentiment analysis," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2022, pp. 419–426.

[55] J. Kocoń, J. Baran, M. Gruza, A. Janz, M. Kajstura, P. Kazienko, W. Korczyński, P. Miłkowski, M. Piasecki, and J. Szołomicka, "Neuro-symbolic models for sentiment analysis," in *International conference on computational science*. Springer, 2022, pp. 667–681.

[56] P. Kazienko, J. Bielaniewicz, M. Gruza, K. Kanclerz, K. Karanowski, P. Miłkowski, and J. Kocoń, "Human-centered neural reasoning for subjective content processing: Hate speech, emotions, and humor," *Information Fusion*, vol. 94, pp. 43–65, 2023.

[57] E. Cambria, Q. Liu, S. Decherchi, F. Xing, and K. Kwok, "Senticnet 7: A commonsense-based neurosymbolic ai framework for explainable sentiment analysis," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 3829–3839.

[58] T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using bilstm-crf and cnn," *Expert Systems with Applications*, vol. 72, pp. 221–230, 2017.

[59] A. Kumar, S. Abirami, T. E. Trueman, and E. Cambria, "Comment toxicity detection via a multichannel convolutional bidirectional gated recurrent unit," *Neurocomputing*, vol. 441, pp. 272–278, 2021.

[60] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[61] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[62] P. Rybak, R. Mroczkowski, J. Tracz, and I. Gawlik, "Klej: comprehensive benchmark for polish language understanding," *arXiv preprint arXiv:2005.00630*, 2020.

[63] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso *et al.*, "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," *Transactions on Machine Learning Research*, 2023.

[64] B. Koptyra, A. Ngo, Ł. Radliński, and J. Kocoń, "Clarin-emo: Training emotion recognition models using human annotation and chatgpt," in *International Conference on Computational Science*. Springer, 2023, pp. 365–379.

[65] K. Kanclerz, A. Figas, M. Gruza, T. Kajdanowicz, J. Kocoń, D. Puchalska, and P. Kazienko, "Controversy and conformity: from generalized to personalized aggressiveness detection," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 5915–5926.

[66] P. Miłkowski, M. Gruza, K. Kanclerz, P. Kazienko, D. Grimling, and J. Kocoń, "Personal bias in prediction of emotions elicited by textual opinions," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, 2021, pp. 248–259.

[67] J. Kocoń, M. Gruza, J. Bielaniewicz, D. Grimling, K. Kanclerz, P. Miłkowski, and P. Kazienko, "Learning personal human biases and representations for subjective tasks in natural language processing," in *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2021, pp. 1168–1173.

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

[69] P. Miłkowski, S. Saganowski, M. Gruza, P. Kazienko, M. Piasecki, and J. Kocoń, "Multitask personalized recognition of emotions evoked by textual content," in *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 2022, pp. 347–352.

[70] K. Kanclerz, M. Gruza, K. Karanowski, J. Bielaniewicz, P. Miłkowski, J. Kocoń, and P. Kazienko, "What if ground truth is subjective? personalized deep neural hate speech detection," in *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, 2022, pp. 37–45.

[71] J. Bielaniewicz, K. Kanclerz, P. Miłkowski, M. Gruza, K. Karanowski, P. Kazienko, and J. Kocoń, "Deep-sheep: Sense of humor extraction from embeddings in the personalized context," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2022, pp. 967–974.

[72] A. Ngo, A. Candri, T. Ferdinan, J. Kocoń, and W. Korczynski, "Studemo: A non-aggregated review dataset for personalized emotion recognition," in *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, 2022, pp. 46–55.

[73] T. Ferdinan and J. Kocoń, "Personalized models resistant to malicious attacks for human-centered trusted ai," in *The AAAI-23 Workshop on Artificial Intelligence Safety (SafeAI 2023)*. CEUR Workshop Proceedings, 2023.

[74] W. Mieleszczenko-Kowszewicz, K. Kanclerz, J. Bielaniewicz, M. Oleksy, M. Gruza, S. Woźniak, E. Dzięcioł, P. Kazienko, and J. Kocoń, "Capturing human perspectives in nlp: Questionnaires, annotations, and biases," in *The ECAI 2023 2nd Workshop on Perspectivist Approaches to NLP*. CEUR Workshop Proceedings, 2023.

[75] J. Kocoń, J. Baran, K. Kanclerz, M. Kajstura, and P. Kazienko, "Differential dataset cartography: Explainable artificial intelligence in comparative personalized sentiment analysis," in *International Conference on Computational Science*. Springer, 2023, pp. 148–162.

[76] W. Wang, H. Bao, S. Huang, L. Dong, and F. Wei, "Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers," *arXiv preprint arXiv:2012.15828*, 2020.

[77] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[78] S. Mukherjee and A. Awadallah, "Xtremedistil: Multi-stage distillation for massive multilingual models," *arXiv preprint arXiv:2004.05686*, 2020.

[79] S. Mukherjee, A. H. Awadallah, and J. Gao, "Xtremedistiltransformers: Task transfer for task-agnostic distillation," *arXiv preprint arXiv:2106.04563*, 2021.