Comprehensive Sentiment Analysis of Polish Book Reviews Using Large and Small Language Models

Agnieszka Karlińska[1], Piotr Miłkowski[2], Paulina Czwordon-Lis[3], Bartłomiej Koptyra[2], and Jan Kocoń[2]
[1]*NASK National Research Institute, Poland*
[2]*Department of Artificial Intelligence, Wroclaw Tech, Poland*
[3]*Institute of Literary Research of the Polish Academy of Sciences, Poland*
agnieszka.karlinska@nask.pl, piotr.milkowski@pwr.edu.pl, paulina.czwordon-lis@ibl.waw.pl,
bartlomiej.koptyra@pwr.edu.pl, jan.kocon@pwr.edu.pl

***Abstract*—This paper presents a comprehensive study of sentiment analysis for Polish book reviews through the creation of a novel, manually annotated dataset and the evaluation of various language models. We introduce a detailed sentiment annotation scheme, addressing challenges encountered during the annotation process, and evaluate model performance on sentiment classification at both the sentence and document levels, as well as text type identification. The study compares specialized Polish transformer models, newly developed Polish-specific large language models (LLMs), and leading commercial LLMs, testing both fine-tuning and zero-shot approaches. Results show that fine-tuned, Polish-adapted LLMs significantly outperform both small language models (SLMs) and commercial zero-shot LLMs, underscoring the importance of domain-specific fine-tuning and language adaptation for sentiment analysis in specialized contexts like literary criticism.**

***Index Terms*—sentiment analysis, book reviews, Large Language Models, Small Language Models, PLLuM**

I. Introduction

Sentiment analysis of book reviews has become an important tool in both literary studies and the publishing industry, providing measurable insights into reader opinions and preferences. This method serves two purposes: it supports academic research into literary trends and cultural shifts, while also offering actionable insights for publishers, authors, and booksellers navigating the ever-evolving book market.

Through sentiment analysis, industry professionals can make data-driven decisions that resonate with their target audience. These include refining marketing strategies, predicting potential bestsellers, and improving manuscript selection processes. Online platforms also benefit from this technology, enhancing recommendation systems that create personalized reading lists based on individual preferences [1]–[3].

In academic research, sentiment analysis aligns with the affective turn in the humanities, which has brought increased attention to the practices of emotional and value-based judgments in literature [4]–[6]. The digital literary landscape is particularly fertile for such research, as the speed and ease of online opinion sharing fosters vibrant literary discussions and intensifies emotional engagement. The role of bloggers and social media in shaping literary reception further highlights the need to study these online interactions to understand contemporary reading practices [7], [8]. Additionally, researchers in scientometrics have applied sentiment analysis to assess the emotional impact of scholarly works, offering a qualitative complement to traditional bibliometric measures [9], [10].

From a bibliographical perspective, the sentiment of reviews holds great value. Documentalists working on literary bibliographies, such as the Polish Literary Bibliography,[1] link review records to their corresponding literary works. While creating literary bibliographies requires deep expertise, there is a growing demand to automate simpler tasks, like identifying book reviews and assigning sentiment.

Datasets for sentiment analysis typically come from product reviews, with Amazon and Goodreads being common sources for book-related analyses (e.g., [1], [3], [9], [11]). These datasets often include some form of annotation, although the extent varies. To date, no fully annotated (i.e., human-labeled) dataset of literary reviews exists for the Polish language. To fill this gap and meet the needs of bibliographers, researchers, and publishers, we have created a manually annotated corpus of Polish book reviews. Literary blogs, rather than cataloging websites or sales platforms, were selected as the data source, given their critical literary content and rich, personal insights from reviewers with diverse styles and genres [7].

This study makes several key contributions:

1) We introduce a high-quality dataset of Polish book reviews, annotated at the sentence and document levels.
2) We develop a detailed sentiment annotation scheme tailored specifically for book reviews, addressing challenges encountered in the annotation process.
3) We conduct a comprehensive evaluation of sentiment analysis models, exploring Small Language Models (SLMs) like HerBERT and Polish RoBERTa, alongside Large Language Models (LLMs) such as the newly developed PLLuM family and Bielik. We also compare these models to commercial LLMs like GPT-4 and Claude 3.5 Sonnet.

Our experiments include fine-tuning both SLMs and LLMs, as well as testing zero-shot capabilities of off-the-shelf

[1]The Polish Literary Bibliography is an annotated bibliography covering literature, theater, and film. It initially spanned 1944/45 to 1988 in printed volumes and is now available online at https://pbl.ibl.poznan.pl/. The bibliography includes records of books, periodicals, performances, and films.

2375-9259/24/$31.00 ©2024 IEEE
DOI 10.1109/ICDMW65004.2024.00064

LLMs. This approach advances sentiment analysis for Polish-language content and provides insights into the effectiveness of language-specific LLM adaptations for specialized NLP tasks. The study contributes to a broader understanding of how to process literary discourse in digital environments.

II. Related Work

Sentiment analysis (SA) of text has gained substantial attention, spanning various domains such as product reviews, movie critiques, and academic texts. However, the sentiment analysis of book reviews, especially in less studied languages such as Polish, remains relatively unexplored. This section reviews key studies on sentiment analysis, focusing on book reviews and methodologies relevant to our research.

A. Sentiment Analysis in Book Reviews

The sentiment analysis of book reviews has been explored using different techniques. Mounika and Saraswathi [12] utilized convolutional neural networks (CNNs) combined with n-grams, showing significant improvements in capturing sentiment polarity in book reviews. Similarly, Srujan et al. [1] employed various machine learning classifiers such as K-Nearest Neighbors (KNN), Decision Trees (DT), Support Vector Machines (SVM), Random Forest (RF), and Naive Bayes (NB) to analyze sentiment in Amazon book reviews, highlighting the effectiveness of feature selection techniques like TF-IDF. The application of supervised classifiers for scholarly book reviews was further demonstrated by Hamdan et al. [13], who addressed domain-specific challenges in sentiment detection.

Other studies have focused on enhancing sentiment analysis through advanced methods and datasets. Almjawel et al. [11] proposed a visual analytics tool to assist users in understanding sentiment trends in Amazon book reviews. Works [14], [15] extended sentiment analysis techniques to less commonly studied languages like Bangla and Hindi, respectively, employing machine learning approaches to achieve high accuracy.

B. Advancements in Sentiment Analysis Techniques

Recent efforts have extended sentiment analysis to multilingual and multidomain contexts. The MultiEmo dataset introduced in [16]–[18] facilitated cross-language validation and demonstrated the versatility of LaBSE embeddings in sentiment classification across multiple languages. This aligns with [19], introducing the PolEmo dataset for Polish multi-domain sentiment analysis, revealing the challenges and opportunities in cross-domain and cross-lingual model adaptations.

Deep learning and transformer models have revolutionized sentiment analysis, particularly for sequential sentence classification tasks. Works [20]–[22] showcased the utility of pretrained models like BERT for contextualized sentence classification, achieving state-of-the-art results without relying on hierarchical encoding. Shang et al. [23] further improved sequential sentence classification with a span-based dynamic local attention model, demonstrating the importance of capturing structural information in the text.

Large Language Models (LLMs) and Small Language Models (SLMs) have emerged as potentially powerful tools for sentiment analysis [24]–[30]. A work [31] demonstrated the effectiveness of ChatGPT in generating synthetic training data to enhance model performance. In contrast, [32] critically evaluated ChatGPT capabilities across various NLP tasks (including sentiment analysis for Polish), providing insights into their strengths and limitations. Transfer learning methods were also explored in [33] for cross-domain learning techniques to improve model performance in sequential sentence classification tasks.

Integrating symbolic knowledge with neural models has shown promise in improving sentiment analysis outcomes [34], [35]. Works [36]–[38] present frameworks like SenticNet and OntoSenticNet, which leverage commonsense knowledge to enhance sentiment analysis, particularly in handling complex semantic dependencies. A work [39] further emphasized the potential of neuro-symbolic approaches by incorporating linguistic knowledge into transformer-based models, yielding significant performance improvements.

C. Conclusion

The current body of research in sentiment analysis reflects diverse methodologies and applications across multiple languages and domains. While significant progress has been made, the sentiment analysis of book reviews remains relatively underexplored. This study aims to bridge this gap by employing classic fill-mask transformers and LLMs, advancing sentiment analysis techniques specifically for Polish book reviews at both document and sentence levels. We aim to comprehensively understand sentiment dynamics in this unique linguistic context by integrating these innovative approaches.

III. Data

This study introduces a novel dataset of reviews sourced from Polish literary and review blogs. These blogs, widely accessible and diverse, serve as a rich online resource for critical literary discourse, offering content ranging from reviews and interviews to author event reports. The abundance of personal impressions and critical evaluations in the blog posts [7] makes them an excellent source for sentiment classification, providing a nuanced and authentic representation of reader responses to literary works.

A. Data Source

The selection of blogs was conducted using an inductive approach. Based on a comprehensive survey of the Polish literary blogosphere carried out by an expert in the field, 120 blogs containing literary reviews were identified. In the next step, blogs containing only fragments of reviews, blogs that could pose particular technical problems during scraping attempts, and four blogs with reviews accompanied by ratings (the latter group was used in the separate automatic annotation task) were removed from the initial list.

The blogs were described with the following metadata: topic (fiction, children's books, specialized literature, diverse topics), number of authors (single- and multi-author blogs), gender of the author(s), proportion of reviews among all texts, and blog type. We created the following typology:

- **Academic review blog:** Run by an author employed in a specific "consecration institution" (e.g., a university), [40] which reinforces their opinion-forming authority. The author writes under their own name, ensuring a high linguistic level. This is the rarest type.
- **Professional review blog:** The author, writing under their own name or occasionally a pseudonym, builds their recognition and brand as a literary commentator. They sometimes strengthen their position by revealing collaborations with magazine editorial teams and/or prominent publishing houses. While care is taken to maintain linguistic quality, many such bloggers may still be criticized for insufficiently professional language and lacking the "editorial-proofreading filter." [41] It should be noted that being a professional, recognizable, popular blogger does not automatically make one a professional literary critic.
- **Amateur (reader's) review blog:** Most commonly written under a pseudonym by readers of various ages and educational backgrounds, dedicated to sharing reading impressions. The linguistic level varies widely. This is the most frequently encountered type.

The blogosphere, as a source of reviews, provides a certain representativeness and diversity, allowing us to fairly consider both more prominent voices and those most numerously represented. The decision to include sources of varying linguistic quality is also inspired by bibliographic practices in the Polish Literary Bibliography, whose creators aim to document all traceable reception of literature, not excluding poorly edited (paper) sources.[2]

The review sources in our collection are also diverse in other aspects: we included both single-author and multi-author blogs, blog-format magazines, and blogs with various specialties and main themes (e.g., new releases, genre literature, selected foreign literature, children's literature, contemporary Polish poetry, contemporary Polish literature, 19th century literature, and selected genres such as reportage, crime fiction, romance, horror literature, and comics).

B. Dataset Preparation

After scraping the content, a total of 48,481 texts from 80 different blogs were obtained. For annotation purposes, a sample of 2,500 texts was randomly selected using the following criteria:

- A minimum of 10 texts from each blog.
- A maximum of 10% of texts from blogs containing content about children's books.
- A maximum of 10% of texts from blogs covering diverse topics.
- A minimum of 15% from multi-author blogs.

We also decided to adjust the proportions in favor of blogs providing reviews with a higher level of editing and linguistic correctness while excluding those that generated technical problems, had too low a linguistic level (number of language errors), contained too much purely commercial content (e.g., price comparison sites), or included few reviews.

The data preprocessing phase involved several crucial steps to prepare the dataset for analysis. Initially, the process focused on cleaning the data, which included removing line breaks and extraneous elements such as URLs, image captions, and other illustrative components. Non-integral parts of the reviews, such as tags and footnotes not containing sentiment-relevant content, were also eliminated. Subsequently, the data underwent sentence segmentation. Various approaches were tested, particularly utilizing libraries such as Stanza [42], Moses [43], and spaCy [44]. Each method generated errors stemming from the diverse formatting of source texts and their linguistic quality (a topic further elaborated in the VII). After careful consideration, the Stanza library was ultimately selected for sentence segmentation due to its overall performance in handling the complexities of the Polish language and the specific challenges presented by the literary blog format.

C. Annotation Scheme

The annotation process was conducted at both the whole-text and sentence levels. Initially, the intention was to annotate only book reviews for sentiment. However, the corpus of texts collected from review blogs contained various types of content, not exclusively book reviews. Due to the lack of metadata and dedicated classification tools, automatic selection was not feasible. Therefore, the first step involved annotating the text type. Based on expert knowledge and the collected material, four labels were identified:

1) **Book review**: Reviews in which the subject of evaluation is a single book (not necessarily literary or related to literary studies), including reviews of poetry collections, comics, anthologies, short story collections, as well as reviews of book series treated as a whole.
2) **Multi-review**: Reviews covering more than one book, excluding series treated as a single entity. These were not sentiment-annotated due to the difficulty in attributing sentiments to specific books within each sentence.
3) **Non-book review**: Reviews of non-book items such as films, exhibitions, or performances.
4) **Non-review**: Texts that do not fit the criteria of a review.

In the second step, texts identified as book reviews underwent sentence-level sentiment annotation. We distinguished three basic labels for sentiment polarity (valence): positive, negative, neutral.

These labels were not mutually exclusive. In cases of ambivalent sentiment, a sentence could be labeled as both positive and negative simultaneously. We also introduced the option to use a "hard to say" label for sentences that were difficult to categorize, such as poorly constructed or incomprehensible statements. Text fragments without substantive content,

[2]This also involves documenting journals that may not necessarily meet the highest editorial standards, including local literary magazines, student publications, and those produced voluntarily or self-published.

such as quotes from the reviewed text, titles, bibliographic descriptions, URLs, and image captions, were excluded from sentiment annotation.

For positive and negative sentiments, we introduced a gradation to indicate intensity (arousal) using two additional labels:

- Weak (slightly positive or negative)
- Strong (strongly positive or negative)

Initially, we assumed that individual sentences should be evaluated independently, without reference to the context of the entire text. However, in practice, this approach was not always appropriate, as it sometimes overlooked the overall sentiment of consecutive sentences that, when analyzed in isolation, did not convey emotion.

The third step involved annotating the sentiment of the entire review by assigning one or more labels: positive, neutral, or negative. Similar to sentence-level annotation, a review could be labeled as both positive and negative. When indicating positive or negative sentiment, annotators also specified the intensity (arousal).

The annotation guidelines, developed by literary scholars and linguists based on domain expertise, material analysis, and pilot annotation results, are described in detail in A. The annotation process followed a 2+1 scheme: each sample was annotated by two annotators, with discrepancies resolved by a super-annotator. Before the main annotation, three rounds of pilot annotation were conducted, each involving 100–200 samples. This led to the refinement of the guidelines, including developing a dictionary of phrases indicating positive or negative sentiment.

To reduce inconsistencies arising from varying levels of expertise among annotators, we established a guideline: contextual information (common knowledge) should only influence annotations when essential for comprehending the text. Annotators were instructed to minimize the use of specialized literary knowledge—defined as expertise in literary studies beyond the high school level—in their assessment process. This approach aimed to ensure a more standardized annotation procedure, relying primarily on widely accessible knowledge rather than advanced literary scholarship. Moreover, regular team meetings were held throughout all annotation stages to resolve doubts and minimize potential discrepancies. Examples of sentence-level sentiment annotations are presented in Table I.

D. Annotation Quality

The inter-annotator agreement was assessed using the Positive Specific Agreement (PSA) measure across various annotation tasks and annotator pairs (Tab. II). The annotation team consisted of five members: two linguists (ling1 and ling2), two literary scholars involved in bibliographic work (lit1 and lit2), and a bibliographer (bibl).

Overall, agreement levels were high for most categories, with some variations across different tasks and annotator pairs. The highest agreement was observed for "book review" classification (95.03% overall), indicating strong consensus in identifying book reviews. "Non-review" classification also showed high agreement (84.06% overall). However, "multi-review" and "non-book review" categories had lower agreement (53.81% and 65.45% respectively), suggesting that these categories were more challenging to distinguish.

Regarding sentiment polarity, neutral sentiment had the highest agreement (89.35% overall), possibly due to the prevalence of neutral statements in reviews. Positive sentiment showed good agreement (78.45% overall), while negative sentiment was slightly lower but still substantial (74.52% overall). Agreement on sentiment intensity (weak vs. strong) was generally lower than on polarity. Strong positive sentiment had higher agreement (69.76% overall) compared to weak positive (53.33% overall). For the negative sentiment, the agreement was similar for both weak (53.33% overall) and strong (53.96% overall) intensities. Interestingly, no annotator pair performed significantly worse than others, but some patterns emerged. Literary scholars with extensive experience in creating bibliographies introduced some variability, as evidenced by lower PSA scores for pairs involving literary scholars in the "book review" category. This may be attributed to their broader perspective on text types, informed by their bibliographic experience, which sometimes blurred the lines between reviews and articles about literary works. However, individual experiences, such as active participation in current literary life, seemed to have a more significant impact on annotation differences than educational background. Age-related biases were also observed, particularly in the lit1-ling2 pair.

The lower agreement scores for "strong negative sentiment," "multi-review," and "non-book review" categories can likely be attributed, in part, to the smaller sample sizes for these types. In these cases, disagreement on a single annotation had a more substantial impact on the PSA score compared to other annotation tasks with larger sample sizes.

The "hard to say" category had a very low agreement (2.08% overall), but this is primarily due to its infrequent use. It was labeled in only six samples in total, which explains the apparent inconsistencies. This low frequency suggests that annotators generally felt confident in their classifications, resorting to "hard to say" only in rare, particularly ambiguous cases.

In conclusion, the inter-annotator agreement analysis reveals strong consensus in identifying book reviews and classifying sentiment polarity, particularly for neutral and positive sentiments. The lower agreement on sentiment intensity and some specific text types highlights areas where the annotation task was more challenging. These findings suggest that while the dataset is reliable for sentiment polarity analysis, caution should be exercised when considering arousal or distinguishing between certain text types. Due to significantly lower and unsatisfactory inter-annotator agreement on sentiment intensity, it was decided to include only annotations related to sentiment polarity in the final dataset.

E. Dataset Statistics

The dataset consisted of 2,500 texts from 80 different blogs, with an average of 31.25 texts per blog (median 20, standard

TABLE I
EXAMPLES OF SENTENCE-LEVEL SENTIMENT ANNOTATIONS IN THE DATASET

Original sentence in Polish	**English translation**	**Sentiment**
Jestem pod olbrzymim wrażeniem, bo nie spodziewałam się tak magnetycznego tekstu.	I am extremely impressed because I did not expect such a magnetic text.	Strong positive
Ostatecznie, choć nie można powiedzieć, że będzie to przełomowa książka na półkach fantastyki dla młodzieży, to zdecydowanie mogę stwierdzić, że jest to książka inna i warta przeczytania.	Ultimately, while one cannot say this will be a ground-breaking book on the young adult fantasy shelves, I can definitely say it's different and worth reading.	Weak positive
Opowieść jest prowadzona w tekście i w ilustracjach kreskówkowych – to książka w takim samym stopniu do czytania, jak i do oglądania.	The story is told through text and cartoon illustrations—it's a book meant equally for reading and for viewing.	Neutral
Historia mnie ani nie przygnębiła (a podejrzewam, że według zamysłu autora choć trochę powinna), ani nie oczarowała literackim kunsztem czy charakterem.	The story neither saddened me (and I suspect it should have, according to the author's intent), nor enchanted me with literary craftsmanship or character.	Weak negative
Okazało się jednak, że powieść jest zbyt przekombinowana, przez co po prostu w pewnym momencie staje się irytująca i nudna.	It turned out that the novel is too overcomplicated, which at some point simply makes it irritating and boring.	Strong negative
Na początku może wydawać się to frustrujące, bo nie wiemy, o co chodzi, ale z każdą kolejną stroną i puzzlem, który wskakuje na swoje miejsce, dostrzegamy piękno tej opowieści.	At first, it may seem frustrating because we don't know what's going on, but with each subsequent page and puzzle piece falling into place, we begin to see the beauty of the story.	Strong positive & weak negative
No, ale niestety wszystko skończyło się na nadziejach, bo autorka sknociła świetnie zapowiadającą się historię.	But unfortunately, it all ended in dashed hopes, as the author botched a story that had such great promise.	Strong positive & strong negative
Przez to książka stała się dla mnie dość przewidywalna i bardzo szybko połapałam się, o co chodzi, ale nie ukrywam, że w pewnym momentach udało się autorce mnie zaskoczyć.	Because of that, the book became quite predictable for me, and I quickly figured out what was going on, but I admit that at certain moments the author managed to surprise me.	Weak positive & weak negative
I naprawdę wciągnęła i liczyłem na więcej, a niestety im dalej w las tym pomysł na Kentuki zgasł jak niektóre maskotki, rozmył się w trywialności i słabym wykończeniu.	It really grabbed me, and I was expecting more, but unfortunately, the further the story went, the idea behind Kentuki faded away like some of the mascots, dissolving into triviality and poor execution.	Weak positive & strong negative

TABLE II
INTER-ANNOTATOR AGREEMENT
PERCENTAGE VALUE OF PSA – POSITIVE SPECIFIC AGREEMENT

Annotators	Book review	Multi-review	Non-book review	Non-review	Positive sentiment	Weak positive	Strong positive	Neutral sentiment	Negative sentiment	Weak negative	Strong negative	Hard to say
lit1 – ling1	91.27	52.63	73.68	83.48	74.48	47.47	64.68	87.34	64.38	47.47	66.02	0.0
ling1 – ling2	95.97	59.26	71.43	81.1	81.34	57.22	74.22	89.43	79.51	57.22	64.09	11.11
lit2 – ling1	94.52	46.67	72.22	88.18	80.67	54.02	69.44	88.91	78.65	54.02	52.94	0.0
bibl – ling2	96.17	56.41	64.15	81.36	76.43	53.07	70.05	90.66	72.85	53.07	49.09	0.0
bibl – ling1	96.74	57.63	63.83	85.13	78.61	54.98	71.06	90.11	73.7	54.98	52.52	0.0
lit1 – lit2	91.53	50.0	66.67	86.15	79.89	47.75	72.02	89.56	68.97	47.75	57.97	0.0
bibl – lit1	92.31	28.57	0.0	88.14	72.89	38.87	61.63	86.7	68.34	38.87	34.34	0.0
lit1 – ling2	91.34	50.0	42.86	76.36	76.44	53.65	62.15	86.3	68.57	53.65	27.45	0.0
All	95.03	53.81	65.45	84.06	78.45	53.33	69.76	89.35	74.52	53.33	53.96	2.08

deviation 30.64). Just over half (52.40%) were from blogs focused on fiction. The second most represented category was blogs dedicated to specialized literature, accounting for 24.04%. As expected, texts from blogs about children's books and those with diverse topics made up just over 10% (10.02%) and below 10% (9.88%), respectively. No data was available on the themes for the remaining blogs.

A significant majority of the texts came from single-author blogs (81.92%). Among the authors, there was a near-equal distribution between amateur and professional writers (46.40% and 45.92%, respectively). Only a small fraction of the texts were written by academics (4.20%). Data on the type of author was missing for the remaining texts.

Nearly half of the texts were from blogs written exclusively by women (46.60%), while another 13.00% were from multi-author blogs created by both genders. Just over one-third of the texts (34.24%) came from blogs written by men. Data was not available for the remaining blogs.

The vast majority of texts came from blogs where reviews were the predominant content (69.16%). Only 5.56% of the blogs had reviews as a minority of their posts.

During the annotation process, a total of 2,476 documents were analyzed and categorized. The distribution of text types was as follows: 1,569 (63.37%) were classified as book

TABLE III
DATASET LABELS STATISTICS FOR REVIEW TYPE.

Split	Book review	Non-review	Non-book review	Multi-reviews
Train	1,267	590	70	53
Eval	148	80	9	10
Test	154	78	10	7
Total	1,569	748	89	70

TABLE IV
DATASET LABELS STATISTICS FOR SENTIMENT.

Level	Split	Positive	Negative	Neutral	Mixed
Documents	Train	892	73	48	238
Documents	Eval	117	8	2	29
Documents	Test	120	11	7	19
Documents	**Total**	1,129	92	57	286
Sentences	Train	7,344	1,902	31,449	566
Sentences	Eval	994	250	3,839	74
Sentences	Test	940	257	3,889	73
Sentences	**Total**	9,278	2,409	39,177	713

reviews, 748 (30.21%) as non-reviews, 89 (3.59%) as non-book reviews, and 70 (2.83%) as multi-reviews.

Of these, 1,564 book reviews, comprising 51,577 sentences, underwent sentiment annotation. The dataset contained a total of 999,438 tokens, with an average sentence length of 19.38 tokens (median 17, std 11.92).

At the sentence level, the sentiment distribution was: 39,177 (75.96%) neutral, 9,278 (17.99%) positive, 2,409 (4.67%) negative, and 713 (1.38%) mixed (both positive and negative). This distribution reveals a predominance of neutral sentences, which is common in review texts where authors often provide objective descriptions or plot summaries alongside their evaluative comments.

Interestingly, the sentiment distribution at the document level showed a different pattern. Among the 1,564 annotated reviews, 1,129 (72.19%) were labeled as overall positive, 286 (18.29%) as mixed, 92 (5.88%) as negative, and only 57 (3.64%) as neutral. The high proportion of positive reviews at the document level may reflect a tendency among reviewers to focus on books they enjoy or a general positivity bias in the literary blogging community.

Table III and Table IV present detailed statistics of the dataset, including the splits into training, validation, and testing sets, as well as the distribution of sentiment labels across these subsets. The dataset split will be further described in the V.

The book review dataset will be made publicly available for scholarly use in accordance with current Polish law on data sharing and intellectual property rights.

IV. MODELS

In this study, we employ a diverse range of language models to evaluate sentiment analysis performance on Polish book reviews. Our selection includes specialized Polish transformer-based models, newly developed Polish-specific LLMs, and state-of-the-art commercial LLMs, allowing for a comprehensive comparison across different model architectures and training approaches.

A. HerBERT, Polish RoBERTa-v2 and Polish Longformer

Transformer-based models for Polish were selected based on the leaderboard from KLEJ Benchmark [45]. The two best models for multiple downstream tasks were selected: HerBERT and Polish RoBERTa (v2). The former is an analog of BERT trained on the Polish language corpus [46], and the latter is its optimized variant using unigram tokenizer, whole word masking, and utilizing larger vocabulary of 128k entries [47]. In addition, Polish Longformer was selected for testing, initialized with Polish RoBERTa (v2) weights and then fine-tuned on a corpus of long documents in Polish, ranging from 1024 to 4096 tokens. The use of the Longformer allows the full content of the review to be considered in the single classification process.

B. Bielik

The newest open-weights model from SpeakLeash, Bielik-11B-v2, is a model initialized from the Mistral-7B-v0.2 model. It has been expanded with additional parameters and trained on 200 billion tokens for two epochs of training. The instruction version of this model has been finetuned on some manually created instructions, but the training mainly consisted of 20 million synthetic instructions generated by the Mixtral 8x22B model. Finally, it was aligned using a DPO-positive algorithm on 66,000 examples.

C. PLLuM Family of Models

This work tests three models from the PLLuM (Polish Large Language Model) family, and each initialized from a different base LLM.

- PLLuM-Llama3-8b – based on Meta-Llama-3-8B model
- PLLuM-Mistral-Nemo-12b – based on Mistral-Nemo-Instruct-2407 model
- PLLuM-Mixtral-8x7b – based on Mixtral-8x7B-v0.1 model

The training, starting from base models of Llama3 and Mixtral models, was performed on about 180 billion tokens from filtered and cleaned Polish corpora. The PLLum model, starting from an instruct version of Mistral-Nemo, has been trained on a smaller subset of high-quality data instead.

All models have been finetuned on a manually created instruction dataset that reflects Polish characteristics and a small subset of additional task-oriented instructions derived from publicly available train sets of Polish NLP datasets.

D. State of the Art LLM Models

For comparison with LLMs tuned strictly for Polish, we also considered the best commercially available models: Mistral Large [*mistral-large-2407*], OpenAI GPT-4o [*gpt-4o-2024-05-13*] and Anthropic Claude 3.5 Sonnet [*claude-3-5-sonnet-20240620*].

V. EXPERIMENTS

Models were tuned for multiclass classification using a fixed random distribution of the set in the proportions of 80% train, 10% validation, and 10% test sets. This split was shared among all experiments.

TABLE V
ACCURACY AND F1-MACRO IN [%] OF LARGE LANGUAGE MODELS.

Model	Documents		Sentences		Text type	
Name (the applied approach)	Acc.	F-1	Acc.	F-1	Acc.	F-1
PLLuM-Llama3-8b (SFT)	87.90	57.00	92.11	85.54	93.17	73.30
Bielik-11B-v2.2-Instruct (SFT)	88.54	**66.44**	91.59	84.47	92.77	79.23
PLLuM-Mistral-Nemo-12b (SFT)	**91.08**	60.72	91.57	84.40	**96.39**	**87.71**
PLLuM-Mixtral-8x7b (SFT)	89.17	58.36	**92.32**	**86.18**	94.38	75.62
Mistral Large (zero-shot)	58.60	46.69	64.97	55.61	68.70	43.94
GPT-4o (zero-shot)	68.79	53.06	70.69	60.00	87.15	71.41
Claude 3.5 Sonnet (zero-shot)	64.33	48.91	71.46	59.68	85.84	63.77

TABLE VI
ACCURACY AND F1-MACRO IN [%] OF SMALL LANGUAGE MODELS.

Model	Documents		Sentences		Text type	
Fine-tuned on tested tasks	Acc.	F-1	Acc.	F-1	Acc.	F-1
HerBERT Large	81.53	61.38	88.87	78.28	**92.77**	75.11
Polish Longformer Large	81.53	60.33	**88.89**	**78.57**	91.16	**76.26**
Polish RoBERTa Large V2	**87.26**	**68.69**	87.89	77.71	92.37	75.97

A. Small Language Model Finetuning

Supervised fine-tuning (SFT) was done using the Transformers library on each of the models (HerBERT, Polish RoBERTa-v2, and Polish Longformer-4096 in the large variant). The baseline learning hyperparameters assumed ten epochs and early stopping after four validation passes, for which there was no improvement in the F-1 measure.

B. Large Language Model Finetuning

For all models, we performed SFT using a newly initialized linear layer to create predictions, replacing the language modeling head that has been pretrained for token prediction. All LLM trainings used the same hyperparameter setup, for ten epochs, with the only difference being that sentence dataset experiments used four times more GPUs and thus four times the effective batch size. Due to computing time constraints, all the experiments were done only once. This means that small differences between experiments, especially on the smaller datasets, documents, and text types, are unlikely to be significant.

C. Zero-shot LLM Approach

To test off-the-shelf LLM models, we used an in-context prompting approach asking the model to evaluate the sentiment of the text or type of review. We provided a list of available classes, followed by the content of the document or sentence to be evaluated. The model was asked to output in JSON format to eliminate possible over-interpretation and to target the classification task.

VI. RESULTS

The most effective approach to our data has been finetuning the language-adapted LLMs. Table V shows the achieved accuracy and F1-macro metrics of the LLMs. Notably, Bielik has achieved the highest F1-macro on the documents dataset despite not having the highest accuracy. This is because it is the only finetuned LLM that correctly classified any of the seven examples of neutral sentiment in the document test set. On the other hand, PLLuM-Mistral-Nemo has a significantly higher F1-score on the text type dataset and the highest accuracy. The notable difference about this model is that it was finetuned and aligned on unknown data before being language-adapted. It is plausible that this model had seen a similar task before the additional training on the Polish language, which transferred to a better score on recognizing the types of texts. To fairly compare the selected finetuned LLMs, these experiments should be repeated, possibly on many sets of hyperparameters. Nonetheless, overall, finetuning of the LLMs consistently outperformed SLMs in terms of accuracy on all datasets and achieved better F1-macro on the highest sample (sentence) dataset.

Using SMLs has proven more effective than the zero-shot approach to powerful LLM models, as seen in Table VI. However, they deviate significantly from the LLMs taught in Polish and tuned to the task. For small models and their low computational complexity, the experiments were repeated ten times and averaged, but the distribution of results, even for the best sample, ranks below the result achieved by Polish LLMs.

VII. DISCUSSION

Sentiment annotation of literary reviews poses challenges at the intersection of natural language processing, literary analysis, and human interpretation. The task's complexity is clear in the contrast between sentence-level and document-level sentiment: reviewers often use neutral language, but their overall assessment is typically positive or negative. This shows that both levels must be considered to fully capture the sentiment in book reviews. A significant technical challenge was sentence segmentation. Errors such as misidentifying initials as sentence breaks or mishandling punctuation affected coherence and complicated the annotation process. Annotators had to assign the same sentiment to split emotional sentences to maintain consistency. The annotation process also uncovered many evaluative sentences referring to works other than the reviewed one, including references to other books or adaptations, which were annotated for sentiment. Partial sentences, ellipses, and sentence fragments without context were particularly difficult to evaluate; introductory or questioning sentences were marked as neutral, while responses carried the sentiment of the exchange.

Handling literary quotations posed another challenge. These were excluded from sentiment annotation to prevent misattribution, though their inconsistent formatting made identification difficult. This underscores the need for better NLP techniques to manage quoted text in future work. Though we aimed to avoid advanced literary knowledge in annotation, understanding the context and genre of reviewed books proved essential. For instance, negative phrases in the context of war literature could reflect a positive assessment of the author's skill. This complexity highlights the necessity of literary competence in the annotation process, but it also introduced biases. Annotators with deeper literary knowledge identified evaluative phrases more effectively [48], while those with stronger linguistic training adhered more closely to dictionary definitions.

Despite these potential biases, the inter-annotator agreement was high for book review identification and sentiment polarity classification, thanks to a rigorous annotation process that included a pilot phase, team meetings, and rotating super-annotator roles. Although challenges remained in sentiment intensity and certain text types, the dataset proved reliable for sentiment polarity analysis.

The experimental results underscore the superior performance of fine-tuned, Polish-adapted LLMs across all tasks, demonstrating the importance of language-specific adaptation. Larger models capture the subtle complexities of sentiment better than smaller models, especially in specialized domains like literary criticism. Although SLMs outperformed zero-shot approaches with commercial LLMs, the gap between models indicates the advantage of fine-tuning on domain-specific data. The poor performance of zero-shot approaches with state-of-the-art commercial LLMs reinforces the need for language- and domain-specific training, particularly in the nuanced literary domain.

VIII. Conclusions and Future Work

This study contributes to sentiment analysis of Polish literary criticism by creating a novel, manually annotated dataset of Polish book reviews and evaluating various language models. We demonstrate that transfer learning from large multilingual models improves performance on language-specific tasks. Poor results from zero-shot learning with commercial models highlight the need for fine-tuning and domain adaptation, underscoring the value of high-quality, domain-specific datasets and fine-tuned models.

The challenges encountered in the annotation process reveal the complexities of sentiment analysis in literary reviews, pointing to areas for future improvement. These include using more advanced NLP techniques for sentence segmentation and quotation detection, considering literary context and genre in sentiment evaluation, and possibly reevaluating the sentence as the primary unit of analysis.

Acknowledgements

This work was financed by (1) the National Science Centre, Poland, project no. 2021/41/B/ST6/04471; (2) CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure (period: 2024-2026) funded by the Polish Minister of Science under the programme: "Support for the participation of Polish scientific teams in international research infrastructure projects", agreement number 2024/WK/01; (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) Digital Research Infrastructure for the Arts and Humanities DARIAH-PL (POIR.04.02.00-00-D006/20-00); (6) Polish Minister of Digital Affairs under a special purpose subsidy No. 1/WI/DBiI/2023: Responsible Development of the open large language model, PLLuM (Polish Large Language Model) aimed at supporting breakthrough technologies in the public and economic sectors, including an open, Polish-language intelligent assistant for public administration clients; (7) 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] K. Srujan, S. Nikhil, H. Raghav Rao, K. Karthik, B. Harish, and H. Keerthi Kumar, “Classification of amazon book reviews based on sentiment analysis,” in *Information Systems Design and Intelligent Applications: Proceedings of Fourth International Conference INDIA 2017*. Springer, 2018, pp. 401–411.

[2] A. Mounika and S. Saraswathi, “Design of book recommendation system using sentiment analysis,” in *Evolutionary Computing and Mobile Sustainable Networks*, V. Suma, N. Bouhmala, and H. Wang, Eds. Singapore: Springer Singapore, 2021, pp. 95–101.

[3] K. Wang, X. Liu, and Y. Han, “Exploring goodreads reviews for book impact assessment,” *Journal of Informetrics*, vol. 13, no. 3, pp. 874–886, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1751157718305078

[4] E. Kim and R. Klinger, “A survey on sentiment and emotion analysis for computational literary studies,” *Zeitschrift für digitale Geisteswissenschaften*, 2019. [Online]. Available: https://zfdg.de/2019_008_v1

[5] S. Ahern, *Affect theory and literary critical practice: a feel for the text*. Cham: Palgrave Macmillan, 2019.

[6] R. Nycz, A. Łebkowska, and A. Dauksza, Eds., *Kultura afektu - afekty w kulturze : humanistyka po zwrocie afektywnym*, ser. Nowa Humanistyka, t. 19. Warszawa: Wydawnictwo Instytutu Badań Literackich PAN, 2015.

[7] A. Nęcka, “Łowcy odsłon. O blogosferze literackiej słów kilka,” *Tematy i Konteksty*, vol. 13, no. 8, pp. 265–276, Dec. 2018, number: 8. [Online]. Available: https://journals.ur.edu.pl/tematyikonteksty/article/view/345

[8] K. Hoffmann, “Hejtuję litblogi? blogi o literaturze a krytyka literacka,” in *Literatura w mediach. Media w literaturze. III. Nowe wizerunki*. Wydawnictwo Naukowe Państwowej Wyższej Szkoły Zawodowej im. Jakuba z Paradyża w Gorzowie Wielkopolskim, 2014, pp. 53–64.

[9] R. Piryani, V. Gupta, V. K. Singh, D. Pinto, D. Pinto, V. K. Singh, A. Villavicencio, P. Mayr-Schlegel, and E. Stamatatos, “Book impact assessment: A quantitative and text-based exploratory analysis,” *J. Intell. Fuzzy Syst.*, vol. 34, no. 5, p. 3101–3110, jan 2018. [Online]. Available: https://doi.org/10.3233/JIFS-169494

[10] Q. Zhou, C. Zhang, S. X. Zhao, and B. Chen, “Measuring book impact based on the multi-granularity online review mining,” *Scientometrics*, vol. 107, no. 3, pp. 1435–1455, Jun. 2016. [Online]. Available: https://doi.org/10.1007/s11192-016-1930-5

[11] A. Almjawel, S. Bayoumi, D. Alshehri, S. Alzahrani, and M. Alotaibi, “Sentiment analysis and visualization of amazon books’ reviews,” in *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*. IEEE, 2019, pp. 1–6.

[12] A. Mounika and S. Saraswathi, “Sentiment analysis of book reviews using cnn with n-grams method,” *International Journal of Knowledge Engineering and Data Mining*, vol. 7, no. 1-2, pp. 64–85, 2021.

[13] H. Hamdan, P. Bellot, and F. Bechet, “Sentiment analysis in scholarly book reviews,” *arXiv preprint arXiv:1603.01595*, 2016.

[14] M. E. Khatun and T. Rabeya, “A machine learning approach for sentiment analysis of book reviews in bangla language,” in *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 2022, pp. 1178–1182.

[15] F. Hussaini, S. Padmaja, and S. Sameen, “Score-based sentiment analysis of book reviews in hindi language,” *International Journal on Natural Language Computing*, vol. 7, no. 5, pp. 115–127, 2018.

[16] 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.

[17] P. Miłkowski, M. Gruza, P. Kazienko, J. Szołomicka, S. Woźniak, and J. Kocoń, "Multi-model analysis of language-agnostic sentiment classification on multiemo data," in *International Conference on Computational Collective Intelligence*. Springer, 2022, pp. 163–175.

[18] ——, "Multiemo: language-agnostic sentiment analysis," in *International Conference on Computational Science*. Springer, 2022, pp. 72–79.

[19] 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.

[20] A. Cohan, I. Beltagy, D. King, B. Dalvi, and D. S. Weld, "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)*, 2019, pp. 3693–3699.

[21] F. Dernoncourt and J. Y. Lee, "Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2017, pp. 308–313.

[22] A. Hassan and A. Mahmood, "Deep learning for sentence classification," in *2017 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*. IEEE, 2017, pp. 1–5.

[23] X. Shang, Q. Ma, Z. Lin, J. Yan, and Z. Chen, "A span-based dynamic local attention model for sequential sentence classification," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021, pp. 198–203.

[24] M. S. U. Miah, M. M. Kabir, T. B. Sarwar, M. Safran, S. Alfarhood, and M. Mridha, "A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm," *Scientific Reports*, vol. 14, no. 1, p. 9603, 2024.

[25] 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.

[26] 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.

[27] 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.

[28] K. Kheiri and H. Karimi, "Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning," *arXiv preprint arXiv:2307.10234*, 2023.

[29] 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.

[30] O. Czeranowska, K. Chlasta, P. Miłkowski, I. Grabowska, J. Kocoń, K. Hwaszcz, J. Wieczorek, and A. Jastrzębowska, "Migrants vs. stayers in the pandemic–a sentiment analysis of twitter content," *Telematics and Informatics Reports*, vol. 10, p. 100059, 2023.

[31] S. Woźniak and J. Kocoń, "From big to small without losing it all: Text augmentation with chatgpt for efficient sentiment analysis," in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2023, pp. 799–808.

[32] 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*, vol. 99, p. 101861, 2023.

[33] A. Brack, E. Entrup, M. Stamatakis, P. Buschermöhle, A. Hoppe, and R. Ewerth, "Sequential sentence classification in research papers using cross-domain multi-task learning," *International Journal on Digital Libraries*, pp. 1–24, 2024.

[34] 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.

[35] J. Kocoń, "Deep emotions across languages: A novel approach for sentiment propagation in multilingual wordnets," in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2023, pp. 744–749.

[36] E. Cambria, "An introduction to concept-level sentiment analysis," in *Advances in Soft Computing and Its Applications: 12th Mexican International Conference on Artificial Intelligence, MICAI 2013, Mexico City, Mexico, November 24-30, 2013, Proceedings, Part II 12*. Springer, 2013, pp. 478–483.

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

[38] M. Dragoni, I. Donadello, and E. Cambria, "Ontosenticnet 2: Enhancing reasoning within sentiment analysis," *IEEE Intelligent Systems*, vol. 37, no. 2, pp. 103–110, 2022.

[39] 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.

[40] D. Antonik, *Sława literacka albo nowe reguły sztuki. Studia z socjologii i ekonomii literatury*. Universitas, 2024.

[41] D. Nowacki, "Bloga se załóż," *Opowiadanie*, no. 2, 2017.

[42] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.

[43] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, S. Ananiadou, Ed. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 177–180. [Online]. Available: https://aclanthology.org/P07-2045

[44] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.

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

[46] R. Mroczkowski, P. Rybak, A. Wróblewska, and I. Gawlik, "Herbert: Efficiently pretrained transformer-based language model for polish," *arXiv preprint arXiv:2105.01735*, 2021.

[47] S. Dadas, M. Perełkiewicz, and R. Poświata, "Pre-training polish transformer-based language models at scale," in *Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part II 19*. Springer, 2020, pp. 301–314.

[48] A. Kałuża, "Jak wytwarza się pojęcia wartościujące? Na przykładzie sporu o poezję Louise Glück," *Forum Poetyki*, no. 28-29, pp. 76–89, Dec. 2022, number: 28-29. [Online]. Available: https://pressto.amu.edu.pl/index.php/fp/article/view/36751

[49] J. Sławiński, Ed., *Słownik terminów literackich*. Zakład Narodowy im. Ossolińskich, 1998.

APPENDIX

A. Annotation Guidelines Summary

This annotation project aims to create training data for a tool that distinguishes literary reviews from other text types and recognizes review sentiment. Annotation occurs at both the whole text level (text type and overall sentiment) and individual sentence level (sentiment).

1) Text Level Annotation: Text Type: First, examine the entire text to determine if it qualifies for further annotation. There are four categories:

1) Book review – a review of a single book (fiction, non-fiction, poetry, comics), anthology, or story collection. Only these will be annotated further.
2) Multi-review – a review of multiple books or multiple volumes of a series.

3) Non-book review – a review of something other than a book, like a film, exhibition, play, or album.
4) Non-review – any other type of text, such as an interview, literary work, event announcement, or article about an author.

Definition of a review: a review is a form of critical commentary. It is an informative-publicistic genre that informs about a cultural fact (including a summary) and serves to express the author's judgment about that fact. According to the Dictionary of Literary Terms, [49], it is primarily an overview, with evaluative elements playing a significant but not dominant role.

For our purposes, we are interested in book reviews. A literary review should contain a clear evaluation of a specific book. Pay attention to the title and first and last sentences for clues. Texts labeled as "preview") are usually not reviews unless they clearly indicate the reviewer has read and evaluated the book. Reviews of book series as a whole can be annotated if they provide an overall assessment, but reviews of individual volumes within a series should be treated as multi-reviews.

2) Sentence Level Annotation: Sentiment: For literary reviews, annotate the sentiment of each sentence as positive, neutral, or negative. Neutral cannot be combined with other labels. Sentences can be both positive and negative if they contain elements indicating both sentiments. Use "hard to say" only when truly uncertain about the sentiment.

Do not annotate sentences that are clearly quotes from the reviewed text, bibliographic information, URLs, image captions, or elements not integral to the review (e.g., "share," "comment"). However, sentences about film adaptations of the reviewed book should be annotated.

For positive and negative sentiments, also indicate the intensity as weak or strong. Consider strategies that strengthen sentiment, such as accumulation of evaluative words/phrases and modifiers/intensifiers. Strong sentiment may be indicated by exclamation marks, capital letters, or discussing multiple aspects in one sentence.

To assist in determining sentiment intensity, a dictionary of words and phrases indicating strong positive or negative sentiment is provided. This dictionary should be used as a reference guide during the annotation process. It includes common expressions and adjectives that typically denote strong sentiment in book reviews.

For your convenience, examples of sentences demonstrating weak positive, strong positive, weak negative, strong negative, and neutral sentiments are provided. These examples serve as a reference point to help calibrate your annotations and ensure consistency across different reviews and annotators.

When annotating, adopt the perspective of the reviewer rather than your own interpretation. If a sentence seems intuitively charged but lacks clear indicators of sentiment, it is better to mark it as neutral. Focus on dictionary meanings rather than personal literary preferences or associations.

3) Text Level Annotation: Overall Sentiment: After annotating individual sentences, determine the overall sentiment of the entire review as positive, neutral, or negative. Reviews can be both positive and negative. For positive or negative sentiment, also indicate the intensity as weak or strong.

Focus on key elements of the review rather than counting individual sentence annotations. Pay particular attention to opening and closing sentences as they often indicate the overall sentiment. Consider the reviewer's overall evaluation of the book, taking into account any concluding remarks or final recommendations.

4) General Remarks: Approach each review and reviewer consistently, regardless of their writing style or tendency towards exaggeration or restraint. Do not automatically assign strong sentiment to a sentence just because it stands out among mostly neutral sentences, and do not default to weak sentiment if there are many strongly charged sentences.

Remember to base your annotations on the text itself, without referring to external sources or applying specialized knowledge beyond general literary understanding. Use the provided dictionary and examples as references to maintain consistency in your annotations.