Behind Closed Words: Creating and Investigating the forePLay Annotated Dataset for Polish Erotic Discourse

Warning: This paper contains examples of sexually explicit material

Anonymous ACL submission

Abstract

The surge in online content has created an urgent demand for robust detection systems, especially in non-English contexts where current tools demonstrate significant limitations. We present forePLay, a novel Polish language dataset for erotic content detection, featuring over 24k annotated sentences with a multidimensional taxonomy encompassing ambiguity, violence, and social unacceptability dimensions. Our comprehensive evaluation demonstrates that specialized Polish language models achieve superior performance compared to multilingual alternatives, with transformer-based architectures showing particular strength in handling imbalanced categories. The dataset and accompanying analysis establish essential frameworks for developing linguistically-aware content moderation systems, while highlighting critical considerations for extending such capabilities to morphologically complex languages.[1]

1 Introduction

The rapid growth of user-generated content online has created an urgent need for effective tools to detect and moderate harmful and inappropriate material. Traditional methods, such as manual review by editors or flagging by users, often fall short due to the sheer volume of content published daily. As a result, there has been an increasing reliance on automated solutions powered by advanced language models and natural language processing techniques.

Securing large language models (LLMs) against the generation of harmful content is another critical concern (Mahomed et al., 2024; Wu et al., 2024). While traditional content moderation tools are useful for user-generated content, they are less effective as input/output guardrails for LLMs due to their inability to adapt to new policies and distinguish between user-generated and AI-generated content (Inan et al., 2023). As a result, there is a growing need for datasets and specialised tools that can be integrated with LLMs to prevent the generation of explicit or harmful material, while maintaining their functionality across different applications.

While much of the previous work has focused on detecting toxicity, abusiveness, offensive language, or hate speech, (Pavlopoulos et al., 2020; Gehman et al., 2020; MacAvaney et al., 2019) these categories do not encompass the full spectrum of undesired content (Markov et al., 2023). Erotic material, in particular, poses a significant risk, especially to minors. Detecting such content is essential not only for creating a safer online environment but also for adhering to ethical standards and legal regulations.

Recent advances in deep learning have improved erotic content detection, but progress is hindered by the limited availability of training data, particularly open datasets. Existing tools are primarily designed for English-language content, limiting their effectiveness for other languages. This linguistic bias highlights the need for language-specific datasets and models that can capture the subtle semantic variations inherent in the expression of erotic content in different languages. Moreover, many existing datasets rely on overly simplified binary classification schemes, which fail to capture the complexities of erotic content, further limiting detection systems' effectiveness in diverse cultural and linguistic contexts.

To address these gaps, we introduce forePLay, the first Polish language manually annotated dataset of erotic content. In contrast to prior efforts, we present a novel multidimensional taxonomy applied to a large Polish language corpus ($n = 24,768$), incorporating fine-grained annotations across ambiguity, violence, and social unacceptability dimensions, while ensuring rep-

[1] The forePLay dataset and associated resources will be made publicly available for research purposes upon publication, in accordance with data sharing regulations.

resentation of LGBT-specific content from both literary and web-based sources. The article provides a detailed description of the annotation process and discusses the challenges encountered during annotation, in particular human label variation. A major contribution is also the comprehensive evaluation of erotic content detection models, examining specialised Polish transformer-based models (HerBERT (Mroczkowski et al., 2021) and Polish RoBERTa (Dadas, 2023)) and Polish-specific LLMs such as the PLLuM family and Bielik (Ociepa et al., 2024). We also compare these models with state-of-the-art multilingual and general-purpose models, including both open-source (Mixtral (Jiang et al., 2024), Llama 3.1 (Dubey et al., 2024), and C4AI Command-R (CohereForAI, 2024)) and commercial solutions (GPT-4o (Hurst et al., 2024)).

2 Related Work

2.1 Existing Datasets

Prior work has established several benchmark datasets for analyzing textual content with erotic themes, though these exhibit notable limitations in scope and annotation granularity. The most comprehensive corpus, Triplex, comprises $27,000$ literary works totaling 1.62 billion tokens extracted from Archive of Our Own (Achour, 2016), while subsequent contributions include the erotic-books[2] corpus ($n = 646$) and the BeaverTails dataset containing a class of *sexually explicit, adult content* and developed for alignment research (Ji et al., 2023). Recent approaches to content analysis have leveraged large language models, as demonstrated in the erotica-analysis dataset[3] ($n = 15,000$) utilizing GPT-3.5 for automated annotation. Various targeted datasets address binary classification of explicit content, including the Jigsaw corpus and its derivatives ($n = 5,100$) (cjadams et al., 2019), though these typically employ simplified taxonomies. While specialized collections such as the sexting corpus[4] ($n = 547$) examine specific discourse patterns, existing datasets predominantly focus on English-language content with binary classification schemes.

[2] https://huggingface.co/datasets/AlekseyKorshuk/erotic-books

[3] https://huggingface.co/datasets/openerotica/erotica-analysis

[4] https://github.com/mathigatti/sexting-dataset

2.2 Erotic Content Detection

Existing methodological approaches to erotic content detection span a broad spectrum of techniques, though their effectiveness is often constrained by the limitations of available training data. Early work relied on classical machine learning approaches, employing Support Vector Machines (SVM) and Random Forest classifiers trained on social media content (Barrientos et al., 2020), while simpler approaches utilized Naive Bayes classification for web page filtering (Hu et al., 2007). More recent neural approaches have demonstrated superior performance, as evidenced by Hierarchical Attention Networks applied to historical Latin texts (n = 2,500 sentences), achieving significant improvements over token-based methods (Clerice, 2024). Contemporary research has increasingly leveraged large language models, as demonstrated by CENSORCHAT (Qiu et al., 2024), which employs knowledge distillation for monitoring dialogue systems. Advanced architectures combining semantic and statistical features through Convolutional Neural Networks (CNN) have shown promise (Wu and Hu, 2005). However, these approaches predominantly focus on English language content, limiting their applicability to other languages where linguistic nuances and cultural contexts play crucial roles in content interpretation.

2.3 Content Moderation Systems

One key application of erotic content detection datasets is content moderation, covering both user-generated and AI-generated material. Specialized systems like Llama Guard, a Llama2-7b safeguard model developed by Meta, use safety taxonomies to categorize prompts, including sexual content (Inan et al., 2023). However, comparative studies indicate that general-purpose LLMs, such as GPT-4o (Hurst et al., 2024) and Gemini 1.5 Pro (Team et al., 2024), often outperform these systems by achieving a better balance between false positives and false negatives (AlDahoul et al., 2024).

Holistic approaches to content moderation, such as those proposed by Markov et al. (2023), combine careful taxonomy design, active learning, and lightweight transformer models. These methods excel in detecting rare harmful content, such as material involving minors. Similarly, the framework introduced by Wu et al. (2024) leverages conceptual features from LLM inference, achieving high accuracy for sexual content with minimal

computational cost. However, their findings reveal that LLMs tend to self-censor compared to human-written text. Similar conclusions were drawn by Mahomed et al. (2024), whose evaluation of OpenAI's moderation systems highlights GPT models' fine-tuning to avoid generating sexual content.

These findings highlight the challenges of building scalable, linguistically inclusive moderation frameworks, especially for nuanced categories like erotic material. Most current tools rely on English-centric training and evaluation pipelines (Markov et al., 2023). To address cultural and linguistic nuances, particularly in morphologically complex languages like Polish, language-specific detection strategies are essential.

3 Data

3.1 Data Collection

We present a large-scale Polish language corpus ($n = 24,768$ sentences) constructed through systematic sampling from two distinct sources: (i) user-generated content from online fiction stories repositories, and (ii) Polish literary works, including Polish translations of world literature. The corpus includes demographically diverse content from 905 unique text units, with a significant manually curated LGBTQ+ representation. Document-level context is preserved through unique file identifiers. While direct source attribution via URLs is omitted due to the volatility of online content, our sampling methodology ensures comprehensive coverage of various linguistic registers and social contexts.

Given the potential application of models in supporting moderation, we focused on internet language and amateur writing, which is predominantly anonymous. Consequently, a significant portion of the sentences (69%) originates from online stories. While the majority of these stories are erotic and non-professional, we included non-erotic stories in a 1:4 ratio to enhance diversity and reduce genre-specific biases. All non-professional stories were scraped from publicly available websites, see Appendix A for details.

To avoid overfitting to specific individual writing styles, we limited the dataset to a maximum of two stories per author. Additionally, incorporating professional literary works with varying degrees of erotic themes (31%) aimed to further diversify linguistic patterns. When selecting these texts, availability in digital form was also a key criterion. A total of 22 different literary texts by Polish and international authors were included. This number is significantly lower than the count of unique text units for non-professional content. However, the literary texts were considerably longer, and finding works that met the established criteria proved challenging. Considering the overall proportion of professional works in the corpus, this number should be sufficient.

3.2 Data Preprocessing

Our text segmentation pipeline employs the NLTK library (Bird and Loper, 2004) for sentence boundary detection, which exhibits robust performance despite the inherent variability in web-sourced content quality. To maximize ecological validity, we preserved the original linguistic characteristics of the source material, including non-standard language patterns and orthographic variations, thereby enabling downstream models to generalize effectively to real-world applications.

The corpus contained a total of 342,546 tokens, as counted using the NLTK tokenizer, with an average sentence length of 13.83 tokens (median 11.0, std 11.53). Detailed token statistics are presented in table 1.

Table 1: Token Count Statistics

Subcorpus	Mean	Std	25%	50%	75%
Total	13.83	11.53	7.0	11.00	17.00
Non-professional	12.43	10.32	6.0	10.00	16.00
Professional	16.90	13.31	9.0	14.00	21.00

Inconsistencies in text segmentation and discrepancies in automatically detected sentences arose from the fact that non-professional writers often did not use proper punctuation, making it challenging for the NLTK library to accurately identify sentence boundaries.

4 Annotation Process

To maintain inclusiveness and diversity in the annotation process, we assembled a gender-balanced team of annotators, comprising 3 male and 3 female individuals, aged 20-40. To minimize potential bias, the annotators were not provided with metadata that could link specific samples to genre types. Additionally, the samples from each genre type were distributed evenly among all annotators. We included annotators with backgrounds in linguistics, literature, or related fields to ensure sensitivity to language nuances. Each sentence was independently assessed by three annotators, with

final labels determined by majority vote. In cases where all three labels differed, a superannotator made the final decision. This occurred in 830 out of $24,768$ samples, representing 3.35% of the total number of observations.

4.1 Annotation Scheme

Based on the expert knowledge of the content, a set of five possible exclusive labels to annotate the samples was designed. This included the following categories: *erotic* (e), *ambigous* (a), *violence-related* (v), *socially unacceptable behaviors* (u), and *neutral* (n).

Sentences labeled as *erotic* typically describe sexual activities and desires, advanced flirting with evident erotic undertones, references to past sexual activities, or explicit sexual fantasies. However, it is important to note that mere mention of genital terms or publicly accepted romantic behaviors, such as kissing or holding hands, does not warrant classification as sexual. Similarly, descriptions of physiological processes or neutral discussions of erotic or sexual topics (e.g., from the psychological perspective) are excluded from this category and considered neutral.

The *violence-related* category is reserved exclusively for sentences that include explicit sexual harassment, rape, lack of consent, or any other sort of non-consensual violence which concerns sexual intentions or activities. It is essential to distinguish consensual BDSM behaviors, as well as non-sexual acts of violence (based solely on the sentence-level analysis), as they do not fall under this category.

The category of *socially unacceptable behaviors* includes sentences that describe sexual behaviors considered illegal, generally taboo, or violative of social norms, such as zoophilia, necrophilia, pedophilia, incest, and other sexual deviations. This category takes precedence over *violence-related* label, meaning that if a sentence could be classified as both *socially unacceptable* and *violence-related*, it should be categorized under *socially unacceptable behaviors*.

Another label covers context-related sentences, which fall under the category of *ambiguous* samples. These sentences are identified based on the conviction that—given commonly recognized patterns of sexual descriptions—they evoke erotic or sexual connotations. However, in a neutral context, they could be interpreted as describing non-sexual behaviors or acts.

Lastly, sentences that did not fall into any of the above-mentioned categories were labeled as *neutral*. On one hand, these typically include non-sexual content on any topic. On the other, this category also comprises general statements about human sexuality or intimate parts that do not typically evoke erotic associations. While such sentences may be sex-related, they are not defined as erotic.

Table 2: Distribution of Label Categories Across Dataset Subcorpora

Subcorpus	**Label Categories**				
	Erotic	Ambiguous	Violence	Unacceptable	Neutral
Total	6,361	1,344	69	116	16,878
Non-professional	2,937	1,277	52	95	12,649
Professional	3,424	67	17	21	4,229

The annotation guidelines, developed by literary scholars and linguists based on domain expertise and material analysis, are described in detail in B.

Following the completion of the superannotation process, the final dataset consists of two primary classes: 68.14% of the samples are labeled as *neutral*, and 25.68% as *erotic*. The remaining categories are minor, with 5.43% labeled as *ambiguous*, 0.47% as *socially unacceptable behaviors*, and 0.28% as *violence-related*. Distribution of labels across the subcorpora are presented in table 2.

The percentage of violence-related examples, as well as those representing socially unacceptable behaviors, was scarce, which was expected, as the data collection process did not specifically target texts exploring such themes. Additionally, while only 15% of the texts in the dataset were sourced from non-erotic stories, the randomization process employed to minimize bias resulted in the inclusion of neutral sentences even from texts classified as erotic, as these are not exclusive to non-erotic stories.

4.2 Annotation Quality

The process of quality evaluation revealed various challenges faced by the annotators, highlighting the demanding nature of the task. The designed labels were intended to be applied exclusively, meaning that each sample could only be categorized as either neutral, sex-related, or ambiguous. Within the sex-related category, overlaps were mitigated by a clear hierarchical structure. As previously mentioned, in cases where a sample exhibited characteristics of both violence and socially unacceptable behaviors, priority was given to the latter. All the other sex-related samples contributed to the erotic class.

Table 3: Examples of annotations in the dataset

Original sentence in Polish	English translation	Label
"Mój mężczyzna..."— usłyszałem szept i zanim się zorientowałem zostałem bez koszulki, a Ty siedziałaś naprzeciw mnie, całując zachłannie moje usta i błądząc dłońmi po torsie, lekko go drapiąc.	"My man..."— I heard a whisper and before I knew it I was left shirtless and you were sitting across from me, kissing my lips greedily and wandering your hands over my torso, lightly scratching it.	erotic
Był cały twardy.	He was all hard.	ambiguous
Ten gad ją zdradzał!	This rat was cheating on her!	neutral

Due to the distinctiveness of rare cases of sexual violence and non-acceptable sexual behaviors, this aspect did not pose significant concerns. Annotators applied these two labels only incidentally, with individual frequencies ranging from 0.2% to 0.6%.

Firstly, the annotation task itself—particularly given the wide range of non-normative language patterns found in non-professional short stories—was inherently subjective. Secondly, the prevalence of misspellings, controversial vocabulary, and grammatical errors, which diminished the clarity of the authors' intentions and made rational decision-making exceedingly difficult. Thirdly, the category of erotic sentences turned out to be extremely broad, spanning from evident and explicit samples, often including vulgar language, to much subtler erotic tones. Consequently, the *ambiguous* label, designed to account for context-related uncertainty, offered a means of resolving indecision between *sexual* and *neutral* classifications. This created a potential risk of overusing the ambiguous label as a way to mitigate uncertainty in borderline cases.

Human label variation (HLV) (Plank, 2022), inherently present in our task of annotating erotic sentences, raised several questions regarding the dataset design (see the 7 Section). To preserve the authenticity of the annotation process, we chose not to intervene beyond the initial training phase. For the dataset's initial release, we opted to follow the conventional practice of data aggregation featuring a ground truth.

To guarantee annotation reliability, we calculated Cohen's Kappa for each pair of annotators across all categories. Furthermore, we evaluated inter-annotator agreement for each category individually. Detailed results are presented in Table 4.

Agreement in most pairs was consistent, ranging from 0.66 to 0.71, regardless of annotators' gender. However, pairs involving annotator F1 presented a notable deviation, as F1 emerged as an outlier, largely due to an excessive use of the ambiguous label.

Table 4: Cohen's Kappa values for each annotator pair across all categories and for each individual category, broken down by gender (F - female, M - male).

Annotators	all	a	e	u	v
F1 / M1	0.1750	0.0532	0.3602	0.2545	0.1984
F1 / F2	0.1808	0.0370	0.3193	0.3952	0.2911
F1 / F3	0.1415	0.0151	0.2890	0.2740	0.2271
M1 / F2	0.7044	0.3505	0.7513	0.3728	0.4602
F2 / M2	0.6910	0.0833	0.7149	0.5074	0.1696
F2 / M3	0.6610	0.0037	0.6923	0.3973	0.3237
F2 / F3	0.7187	0.2683	0.7728	0.6718	0.4078
M2 / M3	0.7030	0.0368	0.7261	0.4191	0.2481

In addition to pairwise agreement, the overall consistency of the entire sample was assessed using Krippendorff's Alpha. The value was quite low (0.387). However, after excluding the F1 annotation, which significantly deviated from the other annotators, the value increased to 0.716, indicating a satisfactory level of agreement in the annotation process.

5 Experiments

We utilize a diverse set of language models for erotic content detection, ranging from specialized Polish encoder-based models and developed Polish-specific LLMs to state-of-the-art commercial LLMs. This variety enables a comprehensive comparison across different model architectures and training methodologies. Additionally, the translated English prompt for the LLMs is included in the appendix on Figure 1.

5.1 Datasets

The original dataset consists of five distinct label categories, but due to significant class imbalance, additional experiments were conducted with modified label groupings as outlined below:

- **Basic**: A binary classification setup where the dataset was reduced to two classes: *neutral* and *erotic* labels.
- **Core**: A three-class classification setup where the dataset included *neutral*, *erotic*, and *ambiguous* labels.
- **Extended**: A four-class classification setup that expanded the dataset to include *neutral*, *erotic* and *ambiguous* labels, with a merged category combining *violence-related* and *socially unacceptable behaviors*.
- **Full**: The full dataset containing all five original categories.

5.2 Specialized Polish Transformer-based Models

Encoder-based models for Polish were selected based on the KLEJ Benchmark leaderboard (Rybak et al., 2020). Two top-performing models across multiple downstream tasks were chosen: HerBERT and Polish RoBERTa. HerBERT is a Polish-language adaptation of BERT, trained on a Polish corpus, while Polish RoBERTa is an optimized version featuring a unigram tokenizer, whole-word masking, and an expanded vocabulary of 128k entries. In our experiments, both models were fine-tuned on each dataset configuration (**Basic**, **Core**, **Extended**, **Full**).

5.3 Polish Large Language Models

Our evaluation framework encompasses three variants from the PLLuM (Polish Large Language Model) family: **PLLuM-Llama 3.1 8B-Instruct**, **PLLuM-Mistral 12B**, and **PLLuM-Mixtral-8x7B**, adapted from their respective base architectures. These models underwent pre-training on 147B Polish language tokens and subsequent instruction tuning. We additionally employed **Bielik 2.3**, trained on 200B tokens and aligned using DPO-positive on 66,000 examples. To assess adaptation capabilities, we fine-tuned PLLuM-Llama 3.1 8B and PLLuM-Mistral 12B on our classification task. All base models were evaluated under 0-shot, 1-shot, and 5-shot configurations, while fine-tuned variants underwent standard supervised evaluation.

5.4 State of the Art Large Language Models

For comparison with LLMs fine-tuned strictly for Polish, we employed a selection of state-of-the-art multilingual and general-purpose models, including both open-sourced and commercial solutions. Among these, **Mixtral**, characterized by its modular and scalable architecture, was utilized in both the 8x22B and 8x7B configurations, alongside **Mistral 12B**, designed as a lightweight alternative. We also tested **Llama 3.1 70B-Instruct** and its smaller counterpart **Llama 3.1 8B-Instruct**, both fine-tuned for instruction-following tasks, as well as **C4AI Command-R**, developed for command-following applications. Additionally, we incorporated **GPT-4o**, frequently used for tasks requiring generalization across various domains.

6 Results

Table 6 highlights the F1 performance metrics of the HerBERT and RoBERTa models across classification tasks involving different dataset configurations. In binary classification scenarios, both models exhibit robust performance. However, as the number of classes increases, a decline in F1 scores is observed. Notably, RoBERTa, particularly in its Base configuration, demonstrates superior performance relative to HerBERT, with pronounced advantages evident in tasks on the **Basic** and **Extended** datasets.

In the case of LLMs, models from the PLLuM family consistently demonstrate strong performance across varying label counts and shot configurations, with particular emphasis on the PLLuM-Mixtral 8x7B, which achieves outstanding F1 scores across a wide range of label counts, notably reaching 0.921 for **Basic** data and 0.670 for **Core** data in the 5-shot setting. Additionally PLLuM-Llama-3.1-8B (SFT) and PLLuM-Mistral 12B (SFT) show significant improvements when subjected to supervised fine-tuning, resulting in enhanced performance, with the highest F1 scores observed across various label counts.

Fine-tuned LLMs exhibit performance levels comparable to traditional transformer-based architectures when evaluated on datasets labeled as **Base** and **Core**, where all categories are sufficiently represented. This suggests that in scenarios characterized by balanced class distributions, where each class is supported by an adequate number of training examples, the performance differential between these methodological approaches is minimal. Conversely, for datasets labeled as **Extended** and **Core**, which are marked by pronounced class imbalance, models built upon the BERT architecture begin to demonstrate a distinct advantage. These models excel in effectively addressing the challenges posed

Table 5: F1 scores for various models across different dataset configurations and shot numbers.

Model	Base			Core			Extended			Full		
	0-shot	1-shot	5-shot	0-shot	1-shot	5-shot	0-shot	1-shot	5-shot	0-shot	1-shot	5-shot
GPT-4o	0.888	0.873	0.891	0.640	0.605	0.618	0.425	0.410	0.415	0.340	0.334	0.349
C4AI Command-R	0.684	0.730	0.714	0.420	0.486	0.454	0.322	0.374	0.363	0.257	0.292	0.299
Llama 3.1 8B-Instruct	0.789	0.736	0.709	0.562	0.517	0.514	0.388	0.364	0.378	0.325	0.283	0.316
Llama 3.1 70B-Instruct	0.837	0.845	0.846	0.595	0.641	0.630	0.396	0.417	0.408	0.329	0.344	0.364
Mistral 12B	0.872	0.847	0.862	0.58	0.579	0.610	0.423	0.411	0.42	0.354	0.349	0.338
Mixtral 8x7B	0.867	0.87	0.862	0.616	0.622	0.621	0.440	0.451	0.435	0.364	0.364	0.348
Mixtral 8x22B	0.872	0.854	0.833	0.633	0.642	0.610	0.424	0.440	0.406	0.335	0.361	0.344
Bielik-11B-v2.2-Instruct	0.848	0.858	0.868	0.601	0.597	0.607	0.427	0.416	0.423	0.312	0.437	0.48
PLLuM-Llama-3.1-8B	0.641	0.737	0.709	0.425	0.446	0.425	0.301	0.324	0.326	0.314	0.274	0.273
PLLuM-Mistral 12B	0.894	0.842	0.844	0.656	0.610	0.613	0.456	0.441	0.427	0.401	0.359	0.401
PLLuM-Mixtral 8x7B	0.874	0.902	0.921	0.647	0.662	0.670	0.452	0.456	0.426	0.422	0.406	0.422
PLLuM-Llama-3.1-8B (SFT)		**0.947**			0.764			**0.461**			0.435	
PLLuM-Mistral 12B (SFT)		0.946			**0.792**			0.458			**0.488**	

by imbalanced data distributions, exhibiting superior capability in identifying underrepresented categories. This highlights their enhanced adaptability and robustness in handling more complex scenarios involving skewed training data distributions.

In the context of multilingual models trained with a primary emphasis on the English language, the results are generally inferior compared to models specifically adapted for the Polish language. The performance varies depending on the configuration, including the number of labels and the number of shots, with different models excelling under different conditions. However, overall, the Mixtrals and GPT-4o models consistently yielded the most favorable results.

The full results of all evaluated models — including the baseline models, the ones adapted to the Polish language from the PLLuM family, and the ones further fine-tuned on our dataset — are presented in Table 5. A detailed account of refusals and responses that could not be classified into any of the predefined categories is presented in Table 7. Among the evaluated models, Llama 3.1 8B-Instruct recorded the highest number of refusals, reaching a total of 848. In contrast, models that consistently provided clear and categorizable outputs included GPT-4o and Mixtral 8x22B, as well as those additionally fine-tuned on our dataset.

Table 6: F1 scores for HerBERT and RoBERTa models across different dataset versions.

Dataset	HerBERT		RoBERTa	
	Base	Large	Base	Large
Base	0.929	0.939	**0.944**	0.943
Core	0.702	0.738	0.738	**0.748**
Extended	0.693	**0.746**	0.704	0.734
Full	0.632	0.648	**0.707**	0.664

7 Discussion

The task of multi-dimensional eroticity annotation is highly complex and inherently subjective, necessitating further and more advanced reflection on the Human-Language Variation (HLV) framework, which has become an increasingly important open problem in recent years (Plank, 2022; Uma et al., 2021; Pavlick and Kwiatkowski, 2019). The researchers highlight the fact that disagreements in the annotation process should not be dismissed as mere *noise* in the search for a single ground truth. On the contrary, such disagreements can often reflect plausible alternative judgments, suggesting that these annotations may not always be categorical (De Marneffe et al., 2012). However, as demonstrated by our experience evaluating annotation quality, distinguishing between clear errors—where no plausible argument could justify the choice of a particular label—and alternative judgments is a critical issue that warrants deeper concern and further discussion. Within the context of erotic content detection, we believe it is valuable to advance current research by utilizing datasets with multiple plausible labels and exploring the approach of Learning with Disagreements (LeWiDi) (Uma et al., 2021) or Multiple Ground Truth (Anand et al., 2024). This approach could

Table 7: Number of unidentified responses or refusals.

Model	Base			Core			Extended			Full			Total
	0-shot	1-shot	5-shot	0-shot	1-shot	5-shot	0-shot	1-shot	5-shot	0-shot	1-shot	5-shot	
GPT-4o	0	0	0	0	0	0	0	0	0	0	0	0	**0**
C4AI Command-R	9	2	14	9	5	10	7	1	9	4	1	14	95
Llama 3.1 8B-Instruct	42	150	12	38	99	14	134	62	8	66	208	15	848
Llama 3.1 70B-Instruct	8	0	0	33	0	0	46	0	0	49	0	0	136
Mistral 12B	0	0	0	0	0	0	0	0	1	0	0	0	1
Mixtral 8x7B	0	0	2	1	0	2	1	5	7	3	0	4	25
Mixtral 8x22B	0	0	0	0	0	0	0	0	0	0	0	0	**0**
Bielik-11B-v2.3-Instruct	116	1	0	115	2	0	148	1	3	2	5	1	394
PLLuM-Mistral 12B	3	5	1	3	2	3	8	7	0	6	10	5	53
PLLuM-Mixtral 8x7B	0	0	1	0	0	2	0	0	2	0	1	2	8
PLLuM-Llama-3.1-8B	1	4	0	1	4	3	1	3	5	0	4	1	27
PLLuM-Mistral 12B (SFT)		0			0			0			0		**0**
PLLuM-Llama-3.1-8B (SFT)		0			0			0			0		**0**

offer more nuanced insights into the field by comparing the performance of models trained on aggregated annotations versus those trained on multiple possible labels.

Our experimental results demonstrate clear performance patterns across model architectures and linguistic contexts, with specialized Polish language models consistently outperforming multilingual alternatives in erotic content detection. The encoder-based architectures, particularly RoBERTa, achieve robust F1 scores in binary classification ($0.929 - 0.944$), though performance degrades with increasing categorical complexity (dropping to ~ 0.66 for five-class classification). This degradation pattern reveals fundamental challenges in maintaining discriminative power across finer-grained categorical distinctions, particularly in cases requiring subtle cultural and contextual understanding.

The PLLuM family of models exhibits exceptional few-shot learning capabilities, with PLLuM-Mixtral 8x7B achieving F1 scores of 0.921 in 5-shot settings for binary classification, while supervised fine-tuning of PLLuM-Mistral 12B yields the highest overall performance (F1=0.946). However, the diminishing returns observed with increasing model size, coupled with the persistent challenge of handling ambiguous content in imbalanced datasets, suggests that architectural sophistication alone may not overcome the fundamental challenges of erotic content detection in morphologically complex languages. These findings underscore the importance of language-specific model development for sensitive content detection tasks, rather than relying solely on increasingly large multilingual models.

8 Conclusions

This work presents forePLay, the first comprehensive Polish language dataset for erotic content detection, introducing a novel multidimensional taxonomy encompassing ambiguity, violence, and social acceptability dimensions. Through systematic empirical evaluation across multiple model architectures, our analysis demonstrates the superiority of specialized Polish language models over multilingual alternatives, particularly in handling nuanced categorical distinctions. The performance patterns observed in transformer-based architectures and few-shot learning scenarios provide valuable insights for developing language-specific content moderation systems. Our findings underscore the critical importance of culturally and linguistically adapted approaches in content detection tasks, while the presented dataset and evaluation framework establish essential groundwork for future research in multilingual content moderation systems, particularly for morphologically complex languages.

9 Limitations

Our study reveals several important limitations that warrant discussion. While our dataset represents a significant contribution to Polish language resources, it exhibits sampling biases through its predominance of content from online fiction reposi-

tories (69%). This may not fully capture the linguistic diversity of erotic discourse across different domains and registers. The relative underrepresentation of professional literary works (31%) potentially limits our understanding of more sophisticated expressions of erotic content.

A significant methodological limitation emerges from our binary approach to violence and socially unacceptable behaviors. The current taxonomy, while practical for annotation purposes, may oversimplify the complex spectrum of content severity and social acceptability. This simplification could impact the generalizability of our models to real-world content moderation scenarios where more nuanced distinctions are crucial.

The annotation process presents limitations stemming from inherent subjectivity in erotic content interpretation. Despite achieving reasonable inter-annotator agreement (Krippendorff's Alpha of 0.716 after outlier removal), the significant variation in individual annotator interpretations—particularly evident in the use of the *ambiguous* category—suggests potential instability in ground truth labels.

From a technical perspective, our evaluation framework primarily focuses on classification accuracy without deeply examining model behavior on edge cases or adversarial examples. While our study demonstrates the superiority of Polish-specific models over multilingual alternatives, we acknowledge limitations in comparing across model architectures with significantly different parameter counts and training regimes.

These limitations underscore the need for future work addressing dataset diversity, annotation methodology refinement, and more robust evaluation frameworks for erotic content detection systems.

References

Sara Achour. 2016. A data driven analysis framework and erotica writing assistant. In SIGTBD.

Nouar AlDahoul, Myles Joshua Toledo Tan, Harishwar Reddy Kasireddy, and Yasir Zaki. 2024. Advancing content moderation: Evaluating large language models for detecting sensitive content across text, images, and videos. Preprint, arXiv:2411.17123.

Abhishek Anand, Negar Mokhberian, Prathyusha Naresh Kumar, Anweasha Saha, Zihao He, Ashwin Rao, Fred Morstatter, and Kristina Lerman. 2024. Don't blame the data, blame the model: Understanding noise and bias when learning from subjective annotations. arXiv preprint arXiv:2403.04085.

Gonzalo Molpeceres Barrientos, Rocío Alaiz-Rodríguez, Víctor González-Castro, and Andrew C. Parnell. 2020. Machine learning techniques for the detection of inappropriate erotic content in text. International Journal of Computational Intelligence Systems, 13(1):591–603.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In Proceedings of the ACL Interactive Poster and Demonstration Sessions, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

cjadams, Daniel Borkan, inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. 2019. Jigsaw unintended bias in toxicity classification. https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification. Kaggle.

Thibault Clerice. 2024. Detecting sexual content at the sentence level in first millennium Latin texts. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 4772–4783, Torino, Italia. ELRA and ICCL.

CohereForAI. 2024. Command r +.

Sławomir Dadas. 2023. Polish roberta.

Marie-Catherine De Marneffe, Christopher D Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. Computational linguistics, 38(2):301–333.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and et al. Abhishek Kadian. 2024. The llama 3 herd of models. Preprint, arXiv:2407.21783.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. arXiv preprint arXiv:2009.11462.

Weiming Hu, Ou Wu, Zhouyao Chen, Zhouyu Fu, and Steve Maybank. 2007. Recognition of pornographic web pages by classifying texts and images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6):1019–1034.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. arXiv preprint arXiv:2410.21276.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama guard: Llm-based

input-output safeguard for human-ai conversations. Preprint, arXiv:2312.06674.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. arXiv preprint arXiv:2307.04657.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L'elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. ArXiv, abs/2401.04088.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. PloS one, 14(8):e0221152.

Yaaseen Mahomed, Charlie M. Crawford, Sanjana Gautam, Sorelle A. Friedler, and Danaë Metaxa. 2024. Auditing gpt's content moderation guardrails: Can chatgpt write your favorite tv show? In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24, page 660–686, New York, NY, USA. Association for Computing Machinery.

Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. Preprint, arXiv:2208.03274.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Krzysztof Ociepa, Łukasz Flis, Remigiusz Kinas, Adrian Gwoździej, and Krzysztof Wróbel. 2024. Bielik: A family of large language models for the polish language - development, insights, and evaluation.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. Transactions of the Association for Computational Linguistics, 7:677–694.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? arXiv preprint arXiv:2006.00998.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Huachuan Qiu, Shuai Zhang, Hongliang He, Anqi Li, and Zhenzhong Lan. 2024. Facilitating pornographic text detection for open-domain dialogue systems via knowledge distillation of large language models. CoRR, abs/2403.13250.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. Klej: Comprehensive benchmark for polish language understanding. arXiv preprint arXiv:2005.00630.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 338–347, Online. Association for Computational Linguistics.

Jialin Wu, Jiangyi Deng, Shengyuan Pang, Yanjiao Chen, Jiayang Xu, Xinfeng Li, and Wenyuan Xu. 2024. Legilimens: Practical and unified content moderation for large language model services. In Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS '24, page 1151–1165, New York, NY, USA. Association for Computing Machinery.

Ou Wu and Weiming Hu. 2005. Web sensitive text filtering by combining semantics and statistics. In 2005 International Conference on Natural Language Processing and Knowledge Engineering, pages 663–667.

A Data Sources

All non-professional stories were scraped from publicly available websites including:

- *opowiadaniaerotyczne-darmowo.com*
- *sexopowiadania.pl*
- *pornzone.com*
- *anonserek.pl*
- *opowi.pl* (categories: *o życiu, różne, miłosne*)
- *opowiadania.pl*
- *polki.pl*

B Annotation Guidelines Summary

This annotation task aims to create a manually annotated dataset for erotic content detection, enabling the development of automated detection methods and content moderation systems. These systems can assist models in identifying undesired content that may be potentially harmful. Each annotator is provided with a set of samples to label exclusively with one of the five predefined categories. The annotation process is conducted independently, meaning that annotators do not have access to the labels or work of other team members. For the purpose of this task, *a sample* is defined as a single sentence of varying length. The style and spelling of the sentences may vary significantly, and assessing the quality or correctness of the text is not part of the task.

Annotators are instructed to carefully read each sentence and assign one final label to it. They are also encouraged to pause or discontinue the task at any point if they feel uncomfortable with the content being evaluated.

Each sample should be assigned with one of the following labels:

Label *e* (for *erotic*): choose this label if a given sentence describes sexual activities and desires, advanced flirting with evident erotic undertones, references to past sexual activities, or explicit sexual fantasies. Please note that mere mentions of genital terms or publicly accepted romantic behaviors or neutral discussions of erotic or sexual topics (e.g., from the psychological perspective) are excluded from this category.

Label *v* (for *violence-related*): choose this label if a given sample depicts sexual behaviors widely considered illegal, taboo, or violative of societal norms. Examples can include references to zoophilia, necrophilia, pedophilia, incest, or other sexual deviations. Please note that a sentence can include violence which is not sex-related. Such instances are excluded from this category.

Label *u* (for *socially unacceptable behaviors*): choose this label if a given sentence describes sexual behaviors considered illegal, generally taboo, or violative of social norms, such as zoophilia, necrophilia, pedophilia, incest, and other sexual deviations. Please note that if you classify a sample as both *violence related* and featuring *socially unacceptable behaviors*, the priority should be given to the latter. Such instances should be label as *socially unacceptable behaviors*.

Label *a* (for *ambiguous*): choose this label to sentences that may evoke some degree of erotic connotation but whose interpretation depends heavily on the context. If the sentence can reasonably be interpreted as non-erotic in a neutral context, it should be labeled as *ambiguous*.

Label *n* (for *neutral*): choose this label if a given sentence does not meet the criteria of the above mentioned sex-related categories. This typically includes non-sexual content on various topics, as well as general statements about human sexuality or intimate anatomy that do not carry erotic connotations.

C Prompt

Prompt Structure:

You are a classifier, and your task is to analyze the text and assign one of the following labels:

1. Neutral text
2. Erotic text
3. Ambiguous text
4. Socially unacceptable behavior text
5. Violence-related text

Based on the text below, return the results in JSON format, which will contain the keys `"Text"` and `"Prediction"`. The value in the `"Prediction"` field should contain only the name of the class.
Input text:

Figure 1: Classification prompt template used for evaluating language models (translated into English).

D Computational Details

The details provided below correspond to the **Full** version of the dataset.

The encoder-based models were trained for approximately 5 hours and 10 hours, respectively, for the Base and Large versions, using A100 (40GB) GPU with early stopping. The batch size was set to 8 for both model variants. Training of the PLLuM family models on our dataset took approximately 3 hours for the PLLuM-Llama 3.1 8B (SFT) model and 4 hours for the PLLuM-Mistral 12B (SFT) model. These models were trained on a distributed

setup consisting of 2 nodes, each equipped with 4 H100 (96GB) GPUs, ensuring parallel processing across GPUs for enhanced training efficiency.