Towards Harmful Erotic Content Detection through Coreference-Driven Contextual Analysis

Inez Okulska and **Emilia Wiśnios**

{inez.okulska,emilia.wisnios}@nask.pl

NASK National Research Institute

Abstract

Adult content detection still poses a great challenge for automation. Existing classifiers primarily focus on distinguishing between erotic and non-erotic texts. However, they often need more nuance in assessing the potential harm. Unfortunately, the content of this nature falls beyond the reach of generative models due to its potentially harmful nature. Ethical restrictions prohibit large language models (LLMs) from analyzing and classifying harmful erotics, let alone generating them to create synthetic datasets for other neural models. In such instances where data is scarce and challenging, a thorough analysis of the structure of such texts rather than a large model may offer a viable solution. Especially given that harmful erotic narratives, despite appearing similar to harmless ones, usually reveal their harmful nature first through contextual information hidden in the non-sexual parts of the narrative.

This paper introduces a hybrid neural and rule-based context-aware system that leverages coreference resolution to identify harmful contextual cues in erotic content. Collaborating with professional moderators, we compiled a dataset and developed a classifier capable of distinguishing harmful from non-harmful erotic content. Our hybrid model, tested on Polish text, demonstrates a promising accuracy of 84% and a recall of 80%. Models based on RoBERTa and Longformer without explicit usage of coreference chains achieved significantly weaker results, underscoring the importance of coreference resolution in detecting such nuanced content as harmful erotics. This approach also offers the potential for enhanced visual explainability, supporting moderators in evaluating predictions and taking necessary actions to address harmful content.

1 Introduction

The identification of harmful content represents a fundamental application of Natural Language Processing (NLP) methods on the internet. Such harmful content encompasses various forms, including hate speech, offensive material, misinformation, and graphic content. Among these, harmful erotic narratives present a particularly sensitive challenge. However, **general adult content detection models primarily focus on distinguishing between non-erotic and erotic texts without nuances in terms of their potential harm**. This is particularly challenging given that the parts describing sexual encounters often appear quite similar in most narratives. **It is the contextual information and additional details describing the individuals involved in or subjected to these sexual actions that ultimately reveal their harmful nature**.

The sentence *"He made love with her"* is a common example of an sexual-related sentence that falls under adult content classification but is generally harmless. However, if the model can detect that elsewhere in the text, at some distance, there is a hint that the term *'her'* refers to a minor or a sibling, it can then raise awareness among readers or, even better, alert moderators to the potentially harmful nature of the text. **And that distant reference, that subtle hint, is precisely what coreference resolution is designed for – to comprehend the semantic chains throughout the narrative**.

In collaboration with moderators from an institution serving as part of a national Incidence Response Team, our research identified a gray area of harmful erotic content that could gain from a coreference-driven contextual analysis. While not yet illegal in certain jurisdictions, this content has the potential to inflict significant harm, particularly on younger or more vulnerable readers.

In this paper, we propose a hybrid context-aware system, using neural and rule-based components, for harmful content detection. It utilizes the coreference module for Polish spaCy model (Tuora and Kobylinski, 2019) to find interrelations between potentially hamrful contextual cues and sex-related parts. While several erotic content classifiers al-

ready exist, there is a conspicuous absence of harmful erotic content classifiers, particularly in languages like Polish. Although our proposed approach has been tested on Polish text, it can readily adapt to other languages, given the availability of BERT-based models and coreference resolution tools for those languages (which is the case for example for the English language).

The primary challenge in developing such a classifier is the scarcity of data. As this form of narrative content straddles the line between illegality and deviance, sourcing and collecting suitable training data poses a formidable obstacle. Unfortunately, the content of this nature falls beyond the reach of generative models due to its potentially harmful nature. **Ethical restrictions prohibit large language models (LLMs) from analyzing and classifying harmful erotics, let alone generating them to create synthetic datasets for other neural models**.

In our research, we assembled, together with the professional moderators, a modest dataset of 164 text samples, meticulously curated and flagged by them. While this dataset remains insufficient for training a robust classifier, it has proven adequate for extracting the harmful contextual features for the hybrid content classifier.

As a result, **we introduce a hybrid model capable of distinguishing between non-harmful and harmful erotic content by leveraging both sexual content predictions and contextual cues**. Our experiments, using real-life examples assessed by professional moderators, demonstrate the promise of this approach, achieving an accuracy of 84% and a recall of 80%. Furthermore, the model's high potential for explainability, thanks to its hybrid coreference-driven architecture, holds great significance for human moderators. They require this level of understanding to evaluate each prediction, make informed decisions, and ultimately classify the text, potentially taking actions such as removing the content from the web or contacting the authorities.

2 Related Work

The task of coreference resolution for the Polish language has been gaining attention for many years (Ogrodniczuk et al., 2014; Nitoń et al., 2018). In 2022, it was also one the subjects of the Shared Task at CRAC 2022 (Saputa, 2022a). There are free coreference tools available – a rule-based and

a statistical resolution tool[1]. Both utilizing the Polish Coreference Corpus (Ogrodniczuk et al., 2016). Applications to which coreference resolution in Polish has been applied include document summarization (Kopeć, 2019) and information extraction (Kaczmarek and Marcińczuk, 2015). In the German language, it has been used for drama analysis (Pagel and Reiter, 2020), and in English, for the analysis of medical interviews (Uzuner et al., 2012).

The issue of gender bias in the Polish language concerning the detection of coreference chains was also studied (Zhu et al., 2016; Zhao et al., 2018; Kocmi et al., 2020)).

As for the detection of harmful content, the harmfulness of which becomes evident only in the context of the interrelationships between mentions, has not been the subject of research so far. This applies primarily to the Polish language, but also, to the best of our knowledge, to other languages.

Efforts in the field of automated Child Sexual Abuse Material (CSAM) detection have predominantly focused on identifying harmful images and videos (Lee et al., 2020). The use of actual CSAM material for model training is constrained by significant legal and ethical complexities. Consequently, researchers have explored alternative approaches, with an emphasis on metadata and filename detection (Pereira et al., 2021).

While textual CSAM content has garnered relatively less attention, Natural Language Processing (NLP) methods and stylometric techniques, such as author profiling, have been adapted for online child grooming detection (Borj et al., 2023). Emil Fleron, utilizing a dataset of abuse forum connections from the 2017 Freedom Hosting 2 dark net leak, investigated how supervised machine learning, relying solely on text data, can identify posts linked to CSAM distribution (Fleron, 2018). Text mining techniques have also been applied to the examination of medical documentation related to child abuse in the Netherlands (Amrit et al., 2017). In a different context, NLP-based methods have been employed to detect sexual/erotic content in user-generated online texts, aimed at filtering out content inappropriate for minors (Barrientos et al., 2020).

Recent studies have ventured into sentence-level pornographic content detection in Chinese and En-

---

[1] http://zil.ipipan.waw.pl/PolishCoreferenceTools

glish datasets comprising novels and stories (Song et al., 2021). However, these approaches are primarily designed to identify adult content, often neglecting the consideration of its harmful or non-harmful nature, and notably, none of them incorporate methods using coreference resolution.

In conclusion, the automated detection of harmful erotic narratives requires further development and investigation, and the incorporation of a coreference-based method represents a novel contribution to this field.

3 Data Collection and Preprocessing

Our hybrid coreference-driven model for harmful erotic content detection relies on two distinct datasets to facilitate and evaluate its functioning in Polish language: a set of sentences describing sexual encounters, called Sexual Sentences Dataset, and the collection of actual harmful erotic narratives, called Harmful Erotic Full-Text Dataset.

3.1 Sexual Sentences Dataset

The dataset comprises approximately 28000 sentences tokenized with NLTK library and selected for binary classification of sexual content with 5865 for class neutral and 22135 for sexual. We intentionally sampled and shuffled these sentences to obscure any contextual cues between adjacent sentences. This deliberate choice prevents the leakage of contextual information from one sentence to the next. For instance, consider the pair of sentences: "He touched her naked skin in a very intimate way. He could see that she loved it." When presented in an unshuffled narrative, the second sentence often led to false positive label 'sexual' due to prior knowledge, although it does not indicate any sexual activity itself. However, by annotating each sentence individually, such ambiguities were minimized.

Each sentence underwent manual labeling by three human annotators, final label was assigned as a result of majority voting. The percentage agreement was 87%. More details regarding annotation process is presented in the Appendix A.

These sentences were sourced from both non-professional and professional narratives gathered from a diverse array of online sources. These sources included web services specializing in short stories contributed by various authors, spanning categories such as "love," "life," "friendship," and "erotic." We did not perform any additional text preprocessing.

3.2 Harmful Erotic Full-Text Dataset

The second dataset served for the main task of the general detection of harmful erotic narratives, and has been split into training, test and validation set. The first one – encompassing 308 samples, has been used for analysis of coreference structures emerging in this type of narrative texts. The same dataset was used for training baseline models (see Section 7). The test set made of 78 samples was used for evaluating the performance of baseline models (Appendix C). The experiments proving the performance of the presented method in numbers, have been run on 164 previously unseen samples.

The harmful class within this dataset comprises text content collected by automated scrapers commissioned by a legal institution tasked with addressing cyber incidents of this nature. Manual classification of these texts was carried out by professionally trained moderators as part of their daily responsibilities.

In the beginning, the full-text corpus exclusively comprised non-professional narratives categorized by moderators under the CSAM (Child Sexual Abuse Material) classification. However, as we collaborated on developing automated classification algorithms, we identified a 'gray zone' of texts. These texts initially fell outside the strict confines of the CSAM definition but were nonetheless deemed disturbing and deviant by both professional and non-professional annotators, including members of the machine learning team.

As a result, we made the decision to expand the category beyond CSAM to encompass harmful-erotic content. This broader category includes all samples describing sexual relations involving young individuals marked by significant age and authority differentials, such as teacher-student dynamics, as well as various forms of incestuous narratives. This expansion specifically encompasses narratives where the focus lies on the sexual excitement induced by family relationships and/or the innocence of a young person, even extending to scenarios involving cousins.

The gathered text samples were tokenized using the same SpaCy model that was employed during the training of the binary sexual sentence classifier. No additional preprocessing steps were applied either before or after tokenization, ensuring the classifier's robustness in handling the natural online

presence of such content.

4 RoBERTa-Based Sexual Sentence Classifier

The neural part of our hybrid approach relies on the sentence-level sexual sentence transformer classifier. It consists of RoBERTa base[2] and additional linear layers with dropout and ReLU activation build on top of the RoBERTa hidden state. The final version utilizes only the base version of the model, as our initial experiments have shown that bigger model does not improve the classification results significantly.

The Sexual Sentence Dataset (as described in Section 3.1) was divided into train (22400) and test (5600) with similar distributions of the classes in both datasets, being approximately 3:1 (non-sexual:sexual).

In its final architecture the model utilizes three linear layers were (with sizes: 768, 512, 256, 1). Dropout probability was equal 0.2. The model was trained on 5 epochs (effect of early stopping based on the validation loss), using Adam as the optimizer with the learning rate of 1e-5. Table 1 shows the results.

Table 1: Classification Report for validation set. "Prec." = precision, "Rec." = recall, "Sup" = support.

	Prec.	Rec.	F1	Sup.
Non-Sexual	0.96	0.96	0.96	4427
Sexual	0.84	0.83	0.84	1173
Accuracy			0.93	5600
Macro avg	0.90	0.89	0.90	5600
Weighted avg	0.93	0.93	0.93	5600

5 Coreference Resolution for Contextual Analysis

Our choice of the coreference resolution method is driven by the recognition that it is the context or scene within these narratives that typically distinguishes harmless erotica from potentially harmful variants. Detailed descriptions of the actors or objects involved in the sexual actions often reside in separate sentences from those describing the actions themselves. The presented method combines information from sentences classified as sexual,

---

[2] https://huggingface.co/sdadas/polish-roberta-base-v2

i.e., those describing sexual activities, with contextual information linked to individuals mentioned in these sentences through a coreference chain as shown in the Algorithm 1. We have decided to use a coreference model for the Polish language proposed by Saputa (2022b), based on the HerBERT model (Mroczkowski et al., 2021). This is an end-to-end model, conveniently integrated into the Polish spaCy model, that achieved an F1 score of 76.67 in the CRAC Shared Task 2022 (Žabokrtský and Ogrodniczuk, 2022) on the Polish test dataset.

First, each sentence receives prediction regarding whether they contain sexual content or not. For the list of all sexual sentences, position indices in the document are determined, ranging from the first to the last token for each sexual sentence. Subsequently, all detected coreference chains in the text are examined to determine if they contain significant contextual information based on the contextual features described in Section 6. If such elements are identified, mention positions are established and compared with the index ranges for the sexual sentences. This process checks whether a given potentially harmful cue refers to another word that is part of a sentence describing sexual activity. Grammatical dependencies are also examined concerning the verbal phrases in the sexual sentence.

Contextual elements in such chains can contain information directly — the same or synonymous noun forms, personal pronouns, or possessives. A crucial aspect of this analysis is the syntactic relationship, which allows us to determine whether, in the case of multiple references to individuals, they are indeed participating in the activities described in the sexual sentence.

Figure 1 illustrates how contextual cues are distributed, and how the coreference mechanism allows for their identification, thereby distinguishing harmful from non-harmful erotic content, as well as shows the visualization potential of this approach. The provided example describes an erotic situation. The text highlighted in red has received a "sexual" label from the binary neural model because it describes sexual activities. However, these words are inherently neutral:

'*Wtedy on wsadził mi rękę pod bluzkę dotykając moich piersi, nachylił się całując me usta*' ('Then he slid his hand under my shirt, touching my breasts, leaned in to kiss my lips').

However, thanks to coreference resolution, it be-

Algorithm 1 Coreference resolution between harmful contextual features and sex-related sentences

Requires:

doc	The SpaCy Doc type
sentences	List of input sentences
coref_chains	List of doc coreference chains
harm_context_feats	List of contextual features (creating harmful context)
seuxal_model	RoBERTa-based sexual sentence classifier

```
 1: sexual_ids ← [ ]
 2: for sentence ∈ sentences do
 3:     sexual_content ← SEXUAL_MODEL(sentence)
 4:     sexual_sent_id ← SENTENCE POSITION IN DOC
 5:     if sexual_content >= 0.5 then
 6:         sexual_ids.APPEND(sexual_sent_id)
 7:
 8: harm_context_clusters ← [ ]
 9: for chain ∈ coref_chains do
10:     chain_ids ← ALL DOC POSITIONS IN CHAIN
11:     for mention ∈ chain do
12:         if mention ∈ harm_context_feats then
13:             harm_context_ids.APPEND(chain_ids)
14:         if harm_context_ids ∈ sexual_ids then
15:             harm_context_clusters.APPEND(chain)
16:     if harm_context_clusters then
17:         label = harmful
18:     else
19:         label = non-harmful
20: return label
```

comes evident that the *'he'* in the sexual sentence is part of a longer chain, which allows us to decipher that it refers to a "guardian" and simultaneously a "P.E. teacher," indicating a physical education teacher.

On its own, the word *'nauczyciel'* ('teacher') can merely serve as a hint about the profession, which is still insufficient to unequivocally classify the content as harmful. However, the second chain leaves no doubt. The analysis of the syntactic dependency in the highlighted sexual sentence demonstrates that the person who is subjected to the sexual activity *'Wtedy on wsadził mi rękę pod bluzkę'* ('he slipped his hand under *my* shirt') is linked in a single chain with the descriptions *'my* guardian' and *'my* P.E. teacher,' unequivocally suggesting that this person is a student and we are dealing with a harmful sexual relationship between a school teacher and his female student.

Additionally, context elements related to the actors participating in sexual sentences but not directly describing the actors themselves, such as elements of their clothing or body parts, were examined. As described in Section 6, the analysis of the training set revealed the unique presence of certain terms. Therefore, we decided to include

them in the set of potentially harmful contextual clues that are detected in the coreference chain encompassing a sexual sentence.

This also applies to various variations of age representation, where the presence of age-related terms was, as in other cases, linked through syntactic dependencies. If a sexual sentence contains subject or object descriptors of sexual activities that, through coreference chains, connect with another sentence containing age-related terms referring to the same person it is a clear indicator for the text label to be 'harmful'. For example, in the sexual sentence *'He touched me'* and the *'me'* connects with *'I'* in another sentence, which is the subject of *'I was 15 years old'*.

6 Contextual Features for Harmful Erotic Content Detection

The coreference features have been automatically extracted and subjected to domain expert analysis. These features form the basis for the rule-based component of the hybrid model and allow for identifying specific elements of context relevant to the analysis of coreference chains related to the sentences describing sexual activities.

From the training set, all text samples involving sexual actions were selected, leaving two classes – of harmful and non-harmful erotic narratives. The TF-IDF analysis was conducted on both of them separately, and based on it, the most frequently occurring words in their base forms (lemmas) were selected for each class. Only those belonging to the category 'noun' (part of speech) were filtered out from them. Then, among them, only those belonging to the 'person' semantic category were chosen, as the most crucial distinguishing element between non-harmful and harmful erotica is the set of actors participating in or being subject to sexual activities. This list ultimately includes 103 nouns.

One of the most striking examples of differences is the word *'mama'* ('mother'), which appeared 303 times in harmful erotica and only four times in the non-harmful class, and *'syn'* ('son'), which appeared 151 times in harmful erotica and 0 times in the non-harmful class. This set of features mainly includes family members, teachers, and terms for children and young people in official, colloquial, and containing typical spelling errors.

Additionally, analyses of both lexicon distributions showed significant differences in the occurrence of additional elements describing scenery –

Mój wfista patrzał na mnie z niezbyt dobrą miną.. nie zwracałam na niego uwagi, zaczęłam ćwiczyć jak inni. Po lekcji mieliśmy pięć minut przerwy, wtedy nic szczególnego się nie działo. Po dzwonku na kolejny wf znów mieliśmy wyjść na salę, ale mój wychowawca mnie zatrzymał (mieliśmy łączone z drugą klasą więc jego obowiązki przejął drugi nauczyciel), kazał mi iść za nim do pokoju nauczycielskiego.

Gdy już weszliśmy, zaczęło się..

- Dlaczego znów się spóźniłaś, Paulinko?

Ja: Zaspałam..

- Kolejny raz? Wiesz, że jak tak dalej pójdzie to nie dam Ci nawet dwójki z powodu Twoich nieobecności. (...)

Powiedział po czym wstał z krzesła i udał się do drzwi, poszłam za nim kiedy on zamknął je na klucz i odwrócił się.. zesztywniałam. Zbliżył się i popchnął mnie na ścianę.. nie wiedziałam co powiedzieć, wyszeptałam tylko 'nie chcę'.

Wtedy on wsadził mi rękę pod bluzkę dotykając moich piersi, nachylił się całując me usta.

Figure 1: Example of visualization the contextual clues through coreference resolution. Marker colors: 1) red - sentence classified as sexual, 2) yellow - one coreference chain referring to the male person involved in the sexual action, 3) green - coreference chain referring to the subject of the sexual action.

parts of clothing and body parts. In the case of the CSAM class, there was an overrepresentation of female clothing elements in the diminutive form (*'spódniczka'*, 'little skirt', *'staniczek', 'camisole')*, also in a characteristic form of non-professional, affective writing, which is not subject to editing, with spelling errors. Importantly, misspelled words do not undergo automatic lemmatization, so it was essential to include them in their original form, e.g.,: *'soudniczke', 'sludniczke', 'spudnice', 'spudnicze', 'spudniczkach', 'spudniczki'* (there are all misspelled version of the word 'little skirt')

Words describing genitalia, characteristic of the harmful class (and not present in the non-harmful class), are vulgar forms that have undergone morphological diminution. On one hand, there is an "adult" term for intimate body parts suitable for sexual actions. On the other hand, there is some adjustment to underage participants (not suitable for sexual actions!) by using diminutive forms. Such a form (e.g., *'kutasik'*, 'tiny little cock') is not encountered either in reference to adults (in that case, the diminutive would imply at least a derogatory, mocking attitude toward the described body part) or in neutral anatomical descriptions of children, where vulgar synonyms for official terms are not used, at most endearing ones (*'siusiak'*,'wee-wee').

As a result of the analysis of extracted nouns, it also turned out to be worthwhile to introduce a negative rule weakening the probability of prediction for the harmful erotic class based on words characteristic only for non-harmful erotica. In the

harmful class, the words *'mąż'* ('husband') and *'żona'* ('wife') did not appear once (only one occurrence of *'żoneczka'*, 'wifey'). In contrast, in the non-harmful erotic class, *'żona'* and *'mąż'* appear pretty often, 113 and 71, respectively.

Additional features detected in the text relate to age and include numerical, verbal, and verbal-numerical representations, taking into account typical forms of misspelling in terms of punctuation and spelling. The upper age limit, which is flagged, is 17 years.

7 Experiments and Results

Detailed description of used RoBERTa model for sentence classification (first step of our classifier to identify sexual vs non-sexual content) is presented in Section 6. As for the full classification, the results were tested on independent real-life data, which consisted of 34 harmful stories and 130 non-harmful stories.

For the full-text classification of harmful erotic content, we compared our proposed coreference-driven approach with RoBERTa base model[3] trained for 10 and for 20 epochs, the Longformer base[4], and a baseline model. Similar to our proposed model, this baseline relies first on identifying sexual sentences with fine-tuned RoBERTa and then is looking for phrases suggesting harmful

---

[3] https://huggingface.co/sdadas/polish-roberta-base-v2

[4] https://huggingface.co/sdadas/polish-longformer-base-4096

Table 2: Results from evaluation of all models on the test dataset made from 164 stories unseed in the training and validation phase. Details regarding parameters used for RoBERTa model and Longformer model are available in Appendix C.

Model	Recall	Precision	F1	Accuracy
RoBERTa base fine-tuned for 10 epoch	91%	30%	63%	45%
RoBERTa base fine-tuned for 20 epoch	70,5%	65%	68%	88%
Longformer	82%	49%	61%	81%
Coreference-Driven Hybrid Classifier	**80%**	**70,5%**	**75%**	**84%**
Baseline Classifier (without coreference resolution)	14%	100%	24%	77,5%

context (both the baseline and coreference-driven model utilize the same dictionaries and semantic rules as described in Section 6). However, the main difference is that the baseline searches for these phrases exclusively in those sentences that have been identified (predicted) as sexual. In contrast, the coreference-driven model seeks contexts in sentences that are not necessary sexual per se (predicted in the first step as "neutral") but connected to sexual ones through the coreference chains. The results of both models demonstrate that the difference in searching for cues in a direct (without utilizing the coreference chains) and a broader context (with coreference chains) is crucial for capturing harmful content.

As shown in Table 2, the RoBERTa base trained for ten epochs seems to be the best in the hunt for the harmful erotica with its 91% recall. However, its focus on the harmful class let the precision drop to a disturbing rate of 30%, meaning that 70% of all non-harmful erotic stories would have been accused of containing some sort of deviation. The analogous issue can be observed in the case of the Langformer model, which offers high recall (82%) but still a very low precision (49%)).

Thus, for the task of detecting harmful erotica the combination of recall, accuracy, and F1-Score is crucial for evaluating such a model instead of solely focusing on the highest recall. Detecting harmful content is a delicate matter since classifying a text with the "harmful" label may even cause legal actions. Therefore, leveraging high recall with high precision is significant in this case. Longer training of RoBERTa improved the precision significantly, but the recall fell by over 20 percentage points, which shows that this architecture cannot find the right balance for this case.

As already mentioned, one of the main challenges for detecting harmful erotic narratives is the collection of the training dataset. Most probably,

the Longformer or the RoBERTa model could have shown more potential when presented with more training data. However, in this real-world case, it was necessary to find a working solution to overcome the problem of such a scarce dataset that was only possible to gather. Also given the fact that this kind of data cannot be effectively generated to enhance the dataset synthetically.

The coreference-driven model achieves a satisfying recall (of 80%) and accuracy (84%), together with good precision and the best F1-Score (75%). As the results show, the coreference-driven rules enhancing the neural sexual sentence classifier offer very promising alternative to the end-to-end models when the training data is lacking.

The results of full-text classification for the coreference-driven classifier are clearly dependent on the performance of the sexual sentence classifier. **Manual analysis of both false negatives and false positives of the coreference-driven hybrid model reveals that the majority of errors stem from an excessive or inaccurate classification of sentences as sexual.** Only in four cases of false negative, despite the accurate classification of sexual sentences, the syntactic relationships in the text proved to be too complex to unequivocally trace the connections between harmful contextual elements and the sexual actions and correctly label the text as harmful. This led us to believe that the proposed approach is perfectly worthy of consideration and further development, mainly focused on improvements of the RoBERTa-based sexual sentence classifier.

8   Discussion

The main challenge in detecting harmful erotica lies in the fact that **merely identifying sexual sentences is insufficient. What makes a text genuinely harmful often requires a comprehensive reading to extract the information from a**

complex context, including entirely non-sexual sentences. This is why the baseline model that searched for the contextual clues only in the direct sexual sentences failed significantly (with the recall of 14%) in the overall detection of harmful content: the clues are usually located outside the sexual sentences. However, previous experiments have also shown that a simple full-text search for the clues (the keyphrases related to age, occupation, or family relationships) is also not enough. That is because these terms can appear in the text, but in a completely neutral context of world-building, unrelated to the sexual activity itself, an innocent part of the depicted world. **Hence, it is only through analysis using coreference resolution that we can search for and determine the true nature of the contextual clues and their relationships to the sexual content**.

9 Limitations and Future Work

Insufficient Harmful Data Availability. A notable limitation of this study lies in the limited availability of harmful data for thorough analysis. The acquisition and accessibility of datasets containing instances of harmful activities have posed significant challenges. To address this limitation, we intend to expand our data collection efforts in future research endeavors. Increasing the volume of available data is vital for enhancing the comprehensiveness and robustness of our analysis and findings.

The model for sexual sentence classification is far from perfect yet. An essential component of the presented algorithm, despite the critical role of coreference, is the detection of sexual sentences within the text. If the model's prediction is incorrect, it can have a negative impact on the overall assessment of the text as harmful because each coreference chain is invariably linked to sentences classified as sexual. In the subsequent stages of the project, we plan to gather and annotate a more diverse set of data and then improve the quality of this classifier.

Adoption of a Learning-Based Approach Over Rule-Based. The current contextual features were manually selected based on tf-idf analysis. With a larger corpus of available texts, it would be feasible to train a dedicated model capable of automatically detecting the presence of these features within the text. This shift toward a learning-based

approach would enhance the system's adaptability and performance, as it could better capture intricate patterns and nuances within the data.

10 Conclusion

Addressing the demanding yet significant application of coreference resolution to harmful erotic content detection, we offer the following contributions:

1. A first neural model fine-tuned solely for classifying sexual sentences in the Polish language, based on the RoBERTa model and trained on 28000 manually annotated sentences.

2. A hybrid neural and rule-based model for detecting harmful erotic content, which leverages coreference resolution to extract necessary contextual clues. This way, it is capable of effectively distinguishing non-harmful from harmful erotic narratives.

3. A visual explanation method for the model potentially highly beneficial for professional moderators involved in the detection of such texts in their work.

4. Preliminary analysis of the issue of harmful erotica in the Polish language.

References

Chintan Amrit, Tim Paauw, Robin Aly, and Miha Lavric. 2017. Identifying child abuse through text mining and machine learning. *Expert systems with applications*, 88:402–418.

Gonzalo Molpeceres Barrientos, Rocío Alaíz-Rodríguez, Víctor González-Castro, and Andrew C. Parnell. 2020. Machine learning techniques for the

detection of inappropriate erotic content in text. *Int. J. Comput. Intell. Syst.*, 13(1):591–603.

Parisa Rezaee Borj, Kiran B. Raja, and Patrick Bours. 2023. Online grooming detection: A comprehensive survey of child exploitation in chat logs. *Knowl. Based Syst.*, 259:110039.

Emil Fleron. 2018. Automatic classification of text regarding child sexual abusive material.

Adam Kaczmarek and Michał Marcińczuk. 2015. Evaluation of coreference resolution tools for polish from the information extraction perspective. In *The 5th Workshop on Balto-Slavic Natural Language Processing*, pages 24–33.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at wmt 2020. *arXiv preprint arXiv:2010.06018*.

Mateusz Kopeć. 2019. Three-step coreference-based summarizer for polish news texts. *Poznan Studies in Contemporary Linguistics*, 55(2):397–443.

Hee-Eun Lee, Tatiana Ermakova, Vasilis Ververis, and Benjamin Fabian. 2020. Detecting child sexual abuse material: A comprehensive survey. *Forensic Science International: Digital Investigation*, 34:301022.

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.

Bartłomiej Nitoń, Paweł Morawiecki, and Maciej Ogrodniczuk. 2018. Deep neural networks for coreference resolution for polish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Maciej Ogrodniczuk, Katarzyna Glowinska, Mateusz Kopec, Agata Savary, and Magdalena Zawislawska. 2014. *Coreference: annotation, resolution and evaluation in Polish*. Walter de Gruyter GmbH & Co KG.

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2016. Polish coreference corpus. In *Human Language Technology. Challenges for Computer Science and Linguistics: 6th Language and Technology Conference, LTC 2013, Poznań, Poland, December 7-9, 2013. Revised Selected Papers 6*, pages 215–226. Springer.

Janis Pagel and Nils Reiter. 2020. Gerdracor-coref: A coreference corpus for dramatic texts in german. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 55–64.

Mayana Pereira, Rahul Dodhia, Hyrum Anderson, and Richard Brown. 2021. Metadata-based detection of child sexual abuse material.

Karol Saputa. 2022a. Coreference resolution for polish: Improvements within the crac 2022 shared task. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 18–22.

Karol Saputa. 2022b. Coreference resolution for Polish: Improvements within the CRAC 2022 shared task. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 18–22, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Kaisong Song, Yangyang Kang, Wei Gao, Zhe Gao, Changlong Sun, and Xiaozhong Liu. 2021. Evidence aware neural pornographic text identification for child protection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14939–14947. AAAI Press.

Ryszard Tuora and Łukasz Kobylinski. 2019. Integrating polish language tools and resources in spacy. In *Proceedings of PP-RAI 2019 Conference*, pages 210–214.

Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 19(5):786–791.

Zdeněk Žabokrtský and Maciej Ogrodniczuk, editors. 2022. *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*. Association for Computational Linguistics, Gyeongju, Republic of Korea.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Rui Zhu, Krzysztof Janowicz, Bo Yan, and Yingjie Hu. 2016. Which kobani? a case study on the role of spatial statistics and semantics for coreference resolution across gazetteers. In *International conference on GIScience short paper proceedings*, volume 1.

A Annotation Details

In this section, we provide a comprehensive account of the annotation process, including the guidelines used, for classifying sentences as either sexual (1) or non-sexual (0) within the scope of our study. To minimize the potential influence of context, the sentences for annotation were intentionally shuffled. The process involved the participation of three annotators for each sentence, with the final label determined through a majority voting mechanism.

A.1 Annotation Process

Annotators. Three human annotators were engaged for the task of sentence classification, each chosen for their proficiency in the target language and their prior experience with similar annotation tasks. These annotators were selected based on their ability to adhere to the annotation guidelines and their capacity to independently assess the sentences.

Majority Voting. To maintain the robustness and objectivity of the annotation process, we employed a majority voting system. For each shuffled sentence, the three annotators independently assigned a label (1 for sexual or 0 for non-sexual). The final label for each sentence was determined through a majority vote. In instances of a tie, a consensus was reached through discussion among the annotators.

A.2 Annotation Guidelines.

Comprehensive and well-defined annotation guidelines were crucial to achieving consistency and accuracy in the annotation process. The following summarizes the key aspects of the annotation guidelines:

Sexual Sentence Definition. A sexual sentence, as stipulated for this annotation task, is one that includes explicit content related to sexual activities or themes. Sentences depicting ordinary acts of affection such as kissing, holding hands, or hugging should not be classified as sexual. Annotators were instructed to focus on the presence of explicit or graphic language, descriptions of sexual acts, or content intended to discuss or explore sexual arousal as indicative of a sexual sentence.

Ambiguity and Context Independence. Given that sentences were presented without context, annotators were instructed to assess each sentence independently. Ambiguity in the sexual nature of a sentence should be resolved based on the sentence's content alone. Annotators should not make assumptions or rely on contextual information that is not explicitly provided.

Consistency and Objectivity. Annotators were encouraged to maintain a consistent approach throughout the annotation process and to avoid the introduction of personal biases. Classification should be solely based on the content of the sentence and its alignment with the provided definition of a sexual sentence.

Annotator Discussions. In cases of uncertainty or disagreement among annotators, open discussions were encouraged to facilitate consensus and ensure the accuracy of the final label. Annotators were allowed to consult relevant reference materials or seek clarification from the research team to address any doubts.

A.3 Inter-Annotator Agreement

To evaluate the reliability of the annotation process, inter-annotator agreement scores were calculated. These scores provide insights into the consistency among annotators and the overall quality of the annotations, considering that sentences were presented without contextual information. We assessed inter-annotator agreement using three commonly employed metrics: Fleiss' Kappa, Cohen's Kappa, and Percentage Agreement.

Fleiss' Kappa. Fleiss' Kappa is a measure of agreement between multiple annotators when categorizing items into multiple categories. For our task of classifying sentences as sexual (1) or non-sexual (0), Fleiss' Kappa was calculated as 0.79. This indicates substantial agreement among annotators.

Cohen's Kappa. Cohen's Kappa measures the agreement between two annotators. It was used to assess pairwise agreement among our annotators. The average Cohen's Kappa across all pairs of annotators was found to be 0.72, indicating substantial agreement between individual pairs.

Percentage Agreement. Percentage agreement, which measures the proportion of sentences for which all annotators agreed on the same label, was 87%.

A.4 Examples

To illustrate the nature of sentences classified as sexual or non-sexual, we provide the following examples from our dataset along with their corresponding annotations in the Table 6.

B Sources of data

All non-professional stories were scrapped 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*

We utilized maximum of 2 stories from the same author.

C Details regarding training parameters

In this section we present parameters used for training baselines models (Table 7) and classification reports on the baselines models (Tables 3, 4, 5).

Table 3: Classification Raport for validation set for baseline RoBERTa model trained on 10 epochs. Prec. means precision and Rec. means recall.

	Prec.	Rec.	F1	Sup.
Non-harmful	0.96	0.71	0.82	38
Harmful	0.78	0.97	0.87	40
Accuracy			0.85	78
Macro avg	0.87	0.84	0.84	78
Weighted avg	0.87	0.85	0.84	78

Table 4: Classification Raport for validation set for baseline RoBERTa model trained on 20 epochs. Prec. means precision and Rec. means recall.

	Prec.	Rec.	F1	Sup.
Non-harmful	0.75	1.0	0.85	38
Harmful	1.0	0.68	0.81	40
Accuracy			0.83	78
Macro avg	0.87	0.84	0.83	78
Weighted avg	0.88	0.83	0.83	78

Table 5: Classification Raport for validation set for baseline Longformer model. Prec. means precision and Rec. means recall.

	Prec.	Rec.	F1	Sup.
Non-harmful	0.82	0.89	0.85	35
Harmful	0.90	0.84	0.87	43
Accuracy			0.86	78
Macro avg	0.86	0.86	0.86	78
Weighted avg	0.86	0.86	0.86	78

Table 6: Examples of Sentence Classification with Sexual Content (A1 – Annotator 1, A2 - Annotator 2, A3 – Annotator 3) **Warning: This table contains sentences with sexual content. Reader discretion is strongly advised.**

Sentence	Translation	A1	A2	A3	Majority Label
Pocałunek trwał kilka sekund.	The kiss lasted for a few seconds.	0	0	0	0
Obciągnęłam spódnicę i cofnęłam nogę pod stół.	I pulled down my skirt and withdrew my leg under the table.	0	0	1	0
Marcin pomógł jej pozbyć się swoich spodni.	Marcin helped her get rid of her pants.	0	1	0	0
Robił to powoli, z czasem przyspieszył, a ja już nie mogłam.	He did it slowly, and over time he sped up, and I couldn't do it anymore.	0	1	1	1
Wstał, a ja uklęknęłam przed nim, wzięłam znowu do ust.	He stood up, and I knelt in front of him, took to my mouth again.	1	1	1	1
Po chwili przyciągnął Joannę do siebie.	After a while, he pulled Joanna close to him.	0	0	0	0

Table 7: Training parameters for baseline models

Parameter	RoBERTa based 10	RoBERTa base 20	Longformer
learning rate	$1e-5$	$1e-5$	$1e-5$
number of epochs	10	10	10
optimizer	Adam	Adam	Adam
batch size	8	8	8
number of linear layers	3	3	3
dropout probability	0.2	0.2	0.2
activation layer	ReLU	ReLU	ReLU