Towards Harmful Erotic Content Detection through Coreference-Driven Contextual Analysis

Inez Okulska, PhD
Emilia Wiśnios

Linguistic Engineering and Text Analysis Department
NASK National Research Institute

SCIENCE
NASK

NASK is a National Research Institute for **Cybersecurity** and **Artificial Intelligence**

National Civil **Cybersecurity Incidence Response Team (CSIRT)**

Supervised by the Ministry of Digital Affairs

SCIENCE
NASK

Linguistic Engineering and Text Analysis Department

RULES, MATCHERS,
LINGUISTIC JARGON

LARGE LANGUAGE
MODELS
FOR EVERYTHING

You

This content may violate our content policy. If you believe this to be in error, please submit your feedback — your input will aid our research in this area.

ChatGPT

This content may violate our content policy. If you believe this to be in error, please submit your feedback — your input will aid our research in this area.


CHILD SEXUAL ABUSE
AUTHORITY EXPLOITATION
INCEST

SEXUAL CONTENT
DETECTION

That day was very hot and shiny.
I was fourteen.
My teacher came to me and took my hand.
He started to kiss and touch me.
I coudn't say a word.

False Negative

No indication of CSAM
in the sexual sentence

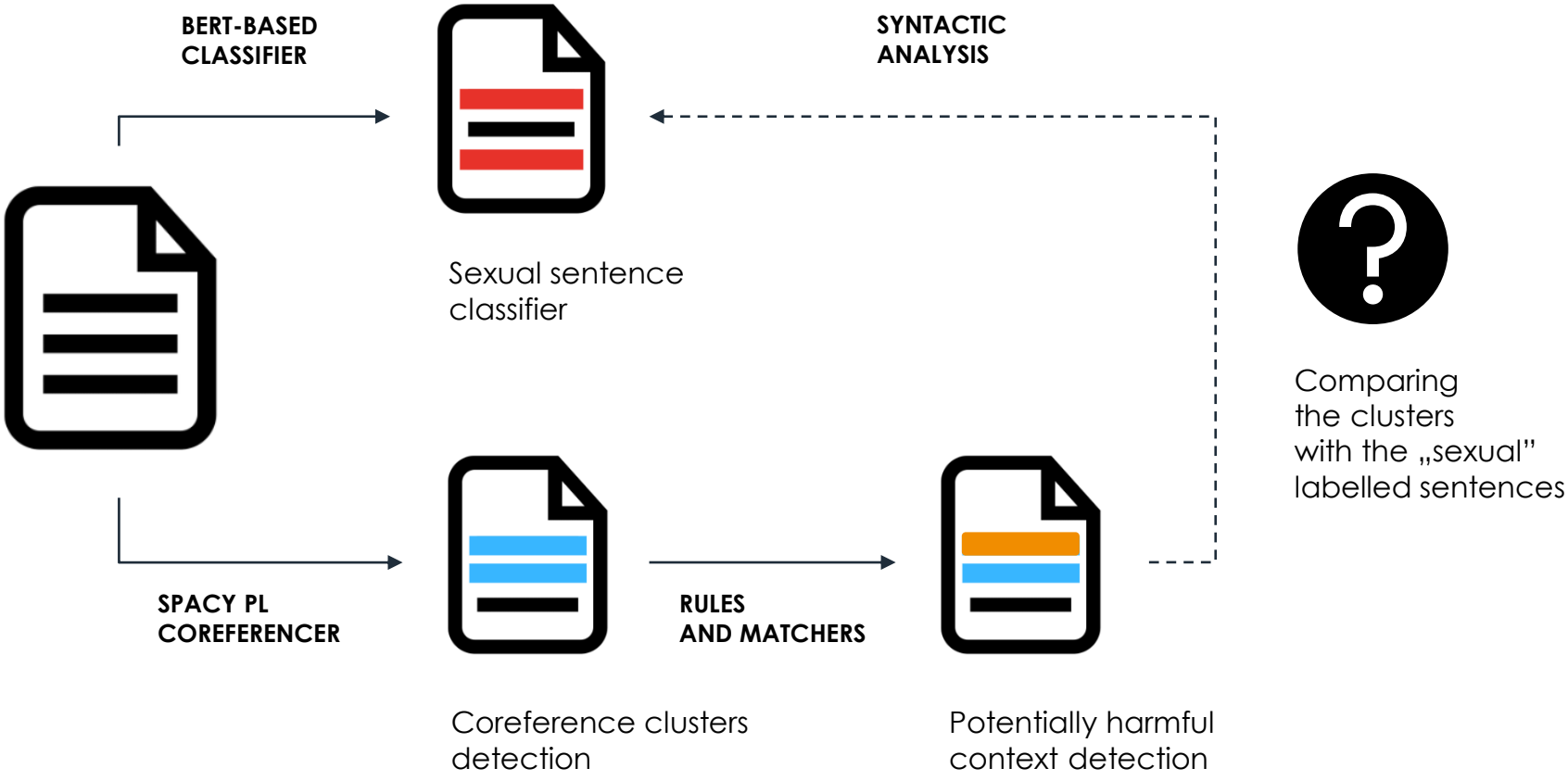
PHRASE SEARCH

That day was very hot and shiny.
I was **fourteen**.
Years later, my **teacher** came to me and took my hand.
I told him about my new boyfriend, how **he started to kiss and touch me**.

False Positive

Contextual phrases detection
without coreference

COREFERENCE-DRIVEN APPROACH


BERT-BASED CLASSIFIER
SYNTACTIC ANALYSIS
Sexual sentence classifier
Comparing the clusters with the „sexual" labelled sentences
SPACY PL COREFERENCER
RULES AND MATCHERS
Coreference clusters detection
Potentially harmful context detection

Sexual sentence classifier

Coreference clusters detection

That day was very hot and shiny.
I was fourteen.
My teacher came to me and took my hand.
He started to kiss and touch me.
I coudn't say a word.

That day was very hot and shiny.
I was fourteen.
My **teacher** came to me and took my hand.
He started to kiss and touch me.
I coudn't say a word.

Cluster 1

A TEACHER started to kiss and touch me.

That day was very hot and shiny.
I was **fourteen**
My teacher came to **me** and took **my** hand.
He started to kiss and touch me
I coudn't say a word.

Cluster 2

CSAM

A TEACHER started to kiss and touch FOURTEEN YEARS OLD me.

TRAINING DATA FOR THE FIRST POLISH SEXUAL SENTENCE CLASSIFIER

938
NARRATIVE TEXTS

Preprocessing:

- Punctuation fixing
- Paragraphs and new lines removal
- Multple white spaces removal
- NLTK-based sentence tokenization

101 949
SENTENCES

28 000
HUMAN ANNOTATED

5865
SEXUAL

22135
NON-SEXUAL

SCIENCE
NASK

RESULTS FOR SEXUAL SENTENCE CLASSIFIER

	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

CONTEXTUAL FEATURES FOR HARMFUL EROTIC CONTENT DETECTION

- School environment indicators
- Age
- Children and teenager-related lexic
- Family members

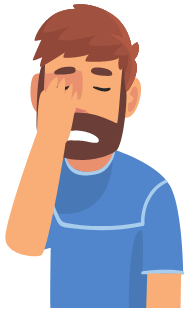
SYNACTIC ANALYSIS OF DETECTED INDICATORS

Based on the dependency to **the object or subject of a sexual activity**

HARMFUL EROTIC DATASET

308
TRAINING SET

78
VALIDATION SET



162
Previously unseen TEST SET

RESULTS OF COREFERENCE-DRIVEN HYBRID CLASSIFIER

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%

SCIENCE
NASK

VISUAL MODEL EXPLANATION WITH COREFERENCE CHAINS

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.

SCIENCE
NASK

CONCLUSIONS

LLM'S are helpless
in the field of harmful content detection

coreference resolution & sentence classifier = **visual explainability**

Nuanced content moderation
can benefit from coreference-driven models

…this is why **further research** of coreference resolution is really crucial and useful!

SCIENCE
NASK

StyloMetrix
NASK

Thank you!

inez.okulska@nask.pl
emilia.wisnios@nask.pl